

McGraw-Hill Series In Education

HAROLD BENJAMIN, *Consulting Editor*

MEASUREMENT IN EDUCATION

McGraw-Hill Series in Education

HAROLD BENJAMIN, *Consulting Editor*

- ALLEN · The Federal Government and Education
BEAUMONT AND MACOMBER · Psychological Factors in Education
BENT AND KRONENBERG · Principles of Secondary Education
BOGUE · The Community College
BROOM, DUNCAN, EMIG, AND STEUBER · Effective Reading Instruction
BRUBACHER · A History of the Problems of Education
BRUBACHER · Modern Philosophies of Education
BUTLER AND WREN · The Teaching of Secondary Mathematics
BUTTERWORTH AND DAWSON · The Modern Rural School
BUTTS · A Cultural History of Education
CARTER AND MCGINNIS · Learning to Read
COOK AND COOK · A Sociological Approach to Education
CROW AND CROW · Mental Hygiene
CROXTON · Science in the Elementary School
DAVIS · Educational Psychology
DAVIS AND NORRIS · Guidance Handbook for Teachers
DE BOER, KAULFERS, AND MILLER · Teaching Secondary English
DE YOUNG · Introduction to American Public Education
FEDDER · Guiding Homeroom and Club Activities
FERNALD · Remedial Techniques in Basic School Subjects
FOREST · Early Years at School
GOOD · Dictionary of Education
HAGMAN · The Administration of American Public Schools
HAMMONDS · Teaching Agriculture
HECK · The Education of Exceptional Children
HOPPOCK · Group Guidance
JORDAN · Measurement in Education
KAULFERS · Modern Language for Modern Schools
McCULLOUGH, STRANG, AND TRAXLER · Problems in the Improvement of
 Reading
McKOWN · Activities in the Elementary School
McKOWN · Home Room Guidance
McKOWN AND ROBERTS · Audio-Visual Aids to Instruction
McNERNEY · The Curriculum

McNERNEY · Educational Supervision
MACOMBER · Teaching in the Modern Secondary School
MAYS · Essentials of Industrial Education
MAYS · Principles and Practices of Vocational Education
MELVIN · General Methods of Teaching
MICHEELS AND KARNES · Measuring Educational Achievement
MILLARD AND HUGGETT · An Introduction to Elementary Education
MORT · Principles of School Administration
MORT AND REUSSER · Public School Finance
MORT AND VINCENT · Modern Educational Practice
MURSELL · Developmental Teaching
MURSELL · Successful Teaching
MYERS · Principles and Techniques of Vocational Guidance
PITTENGER · Local Public School Administration
REMMLEIN · The Law of Local Public School Administration
REMMLEIN · School Law
RICHEY · Planning for Teaching
SAMFORD AND COTTLE · Social Studies in the Secondary School
SANFORD, HAND, AND SPALDING · The Schools and National Security
SCHORLING · Student Teaching
SCHORLING AND WINGO · Elementary-school Student Teaching
SEARS · The Nature of the Administrative Process
SMITH, STANDLEY, AND HUGHES · Junior High School Education
SORENSEN · Psychology in Education
THORPE · Psychological Foundations of Personality
THUT AND GERBERICH · Foundations of Methods for Secondary Schools
TIDYMAN AND BUTTERFIELD · Teaching the Language Arts
WARTERS · High-school Personnel Work Today
WELLS · Elementary Science Education
WELLS · Secondary Science Education
WILSON, STONE, AND DALRYMPLE · Teaching the New Arithmetic
WINSLOW · Art in Elementary Education
WINSLOW · The Integrated School Art Program

Measurement in Education

An Introduction

A. M. JORDAN

Professor of Educational Psychology
University of North Carolina

New York Toronto London
McGRAW-HILL BOOK COMPANY, INC.

1953

YU 577-211-04
271
E

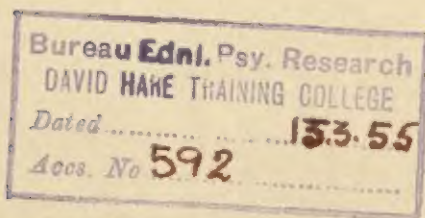
371.26
JOR

MEASUREMENT IN EDUCATION

Copyright, 1953, by the McGraw-Hill Book Company, Inc. Printed in the United States of America. All rights reserved. This book, or parts thereof, may not be reproduced in any form without permission of the publishers.

Library of Congress Catalog Card Number: 52-6540

II



Preface

There are two points of view extant which have influenced the construction of textbooks on measurement in education. One of these develops logically the history and principles of testing. Samples and items of tests are used mainly for illustrating the principles. There is no detailed study of particular tests. A second point of view describes the tests in detail but places little emphasis on test construction or on the more fundamental principles involved in measurement.

The present text may be thought of as resulting from a combination of these two points of view. The thought here is that a great many details are necessary to develop the principles which are present in the test items. In case after case the principles involved in test construction are pointed out to the reader. One fundamental concept, frequently illustrated, is that a test score is merely a sample of an individual's performance. Because students need to discuss some tests in great detail, the critical approach used in this text may make them more sensitive to the principles involved.

Considerable emphasis is given to the testing of reasoning and understanding. Samples of attempts to measure these characteristics are introduced even though the tests are tentative and unavailable. They furnish an earnest of the direction of future testing development.

The influence of my teachers, Edward L. Thorndike, Robert S. Woodworth, and F. N. Freeman can easily be detected in my treatment of measurement. More recently, the publications of the former Progressive Education Association have influenced me greatly. It seems to me that their methods of test construction as exemplified in *Appraising and Recording Student Progress* are sound.

My obligations are many. Publishers of tests have been very kind in permitting the use of items, charts, and graphs which frequently have been taken out of their context. At appropriate places in the text recognition is given. Some of my colleagues have also helped by critically reading parts of the manuscript and furnishing helpful suggestions. William H. Peacock has read the chapter on measurement in physical education; Charles M. Clark, the chapter on the measurement of the

social sciences; Mary Bynum Pierson, the chapter on statistics; and Carl F. Brown, the section on reading. My wife Carrie Nicholson Jordan has read the entire manuscript and contributed much to its clarity of expression and its meaning. My thanks go out to them all.

A. M. JORDAN

CHAPEL HILL, N.C.
August, 1952

Contents

| | | |
|---|--|-----|
| PREFACE | | vii |
| PART ONE. PROBLEMS OF MEASUREMENT | | |
| ✓ 1 INTRODUCTION | | 3 |
| Difficulties in Measuring Mental Traits. Results of Developing Units of Measurement. Measurement in Guidance. Measurement in Education. Summary, Questions and Exercises, Bibliography | | |
| ✓ 2 CHARACTERISTICS OF MEASURING INSTRUMENTS | | 14 |
| Internal Validity. External Validity. Recent Trends in Test Validation. Vitiating Factors in Validity. Reliability. Administrability. Interpretation and Comparability. Economy. Summary, Questions and Exercises, Bibliography | | |
| ✓ 3 CONSTRUCTING ACHIEVEMENT TESTS | | 40 |
| Constructing Classroom Tests. Essay-type Questions. Short-answer Questions. Organization and Arrangement of Tests. Improving the Essay Type of Examination. Summary, Questions and Exercises, Bibliography | | |
| ✓ 4 THE TESTING PROGRAM—ACHIEVEMENT-TEST BATTERIES | | 67 |
| Planning for the Testing Program. Development of Achievement-test Batteries. Summary, Questions and Exercises, Bibliography | | |
| 5 MEASUREMENT OF READING, SPELLING, AND HANDWRITING | | 95 |
| Reading. Spelling. Handwriting. Summary, Questions and Exercises, Bibliography | | |
| 6 MEASUREMENT OF LANGUAGE AND LITERATURE | | 144 |
| Aims and Objectives of Teaching Language. Summary, Questions and Exercises, Bibliography | | |
| 7 MEASUREMENT OF THE SOCIAL SCIENCES | | 183 |
| Objectives in the Teaching of the Social Sciences. Measurement of Objectives. Measurement of Achievement in the Social Studies. Summary, Questions and Exercises, Bibliography | | |

| | | |
|----|--|-----|
| 8 | MEASUREMENT OF FOREIGN LANGUAGES. | 207 |
| | Objectives in Teaching. The More Measurable Objectives. Tests of French. Spanish Tests. German Tests. Italian Tests. Latin Tests. Evaluation of Tests of Foreign Languages. Summary, Questions and Exercises, Bibliography | |
| 9 | MEASUREMENT OF MATHEMATICS | 225 |
| | Importance of Mathematics in Our Modern World. Tests of Mathematics in the Elementary School. Tests of Mathematics in High School. Summary, List of Tests in Mathematics, Questions and Exercises, Bibliography | |
| 10 | MEASUREMENT OF SCIENCE | 248 |
| | Aims and Objectives of Science Teaching. Tests of Science in the Elementary School. Tests of Sciences in High School. Scientific Thinking. Attitudes and Interests in Science. Summary. List of Science Tests. Summary, Questions and Exercises, Bibliography | |
| 11 | MEASUREMENT OF BUSINESS EDUCATION | 273 |
| | Objectives in Business Education. Problems of Testing. Clerical Tests. Tests of Clerical Aptitudes. Clerical Achievement Tests. Bookkeeping Tests. Content Tests. Summary, Questions and Exercises, Bibliography | |
| 12 | MEASUREMENT OF FINE ARTS AND MANUAL ARTS | 288 |
| | Music. Art. Manual Arts. Mechanical Aptitude and Ability. Summary, Questions and Exercises, Bibliography | |
| 13 | MEASUREMENT OF PHYSICAL EDUCATION AND HEALTH | 335 |
| | Objectives in Physical Education. Tests of Physical Capacities. Cardiovascular Tests. Tests of Strength. Tests of Posture. Tests of Motor Coordination. Achievement Tests. Measurement and Health Information. List of Tests of Health Education. Tests of Information in Physical Education. Summary, Questions and Exercises, and Bibliography | |
| | ✓PART TWO. MEASUREMENT OF INTELLIGENCE | |
| 14 | INTELLIGENCE AND ITS MEASUREMENT | 358 |
| | Development of Intelligence Tests. Individual Tests of Intelligence. The Meaning of Intelligence. Summary, Questions and Exercises, Bibliography | |
| 15 | GROUP TESTS OF INTELLIGENCE | 378 |
| | Development of Group Tests. Primary Mental Abilities. Intelligence Tests for Various Levels. Uses of Intelligence Tests. Results of Educational Guidance. Uses of Intelligence Tests in Homogeneous Grouping. Aids in Making Decisions about Going to College. Uses of Intelligence | |

Tests for Vocational Guidance. Summary, Questions and Exercises,
Bibliography

PART THREE. PERSONALITY INVENTORIES

16 MEASUREMENT OF INTEREST 423

Characteristics of Interests. Methods of Discovering Interests. Uses of
Interest Inventories. Summary, Questions and Exercises, Bibliography

17 MEASUREMENT OF ATTITUDES 447

Measurement of Attitudes. Summary, Questions and Exercises, Bibli-
ography

18 MEASUREMENT OF PERSONALITY TRAITS 465

Self-inventories or Questionnaires. Validity of Personality Inventories.
Rating Scales. Summary, Questions and Exercises, Bibliography

PART FOUR. STATISTICAL METHODS

19 STATISTICAL METHODS 499

Assembling the Data. Summary, Questions and Exercises, Bibliography

INDEX 523

PART ONE

Problems of Measurement

CHAPTER 1

Introduction

The process of education includes three major divisions: (1) the determination of goals or objectives, (2) the manipulation of materials and methods so that these objectives are achieved, and (3) the evaluation or appraisal of results obtained. In general, it is the function of philosophy to decide upon and define in terms of pupil or student behavior the outcomes or objectives of education. It is the function of psychology to discover the principles of learning and of the nature of childhood so that the most efficient methods and the most suitable material may be chosen and also so that the objectives may be achieved in the most efficient manner. It is the *function of measurement* to furnish such exact information about the outcomes of education that their evaluation and appraisal can be made with more certainty and with a greater degree of truth.

In the past, it has been assumed that experts were needed for the determination of objectives and the selection and adaptation of methods and materials to the level of achievement reached by the child. There has been much less concern about the examinations, ratings, and other methods of measuring the outcomes of instruction. These latter have all too frequently been evaluated by means of hastily constructed examinations and quizzes or by ratings which not seldom have been influenced by that mixture of many ingredients called the school mark. It is also well known that a judgment of value or appraisal is accurate in proportion as it is based on carefully collected information. From the days of Starch and Elliott¹ who sent around a photostatic copy of a geometry paper to be graded by teachers of mathematics, to Hartog's *Examination of Examinations*,² in which such divergent marks were given to the same examination paper by professional readers of examinations, there have accumulated masses of evidence showing the inadequacy and unreliability of the ordinary essay examination. Yet this form of testing is today perhaps more widely used than any other.

¹ Starch, Daniel, and Edward C. Elliott, "Reliability of Grading High School Work in Mathematics," *School Review* (1913) 21:254-259.

² Hartog, Sir Philip, and E. C. Rhodes, *An Examination of Examinations*. New York: The Macmillan Company, 1935.

It is thus clear that appraisals based on information gained from hastily constructed tests or from subjective impressions of teachers cannot have that element of certainty so necessary in the evaluation of objectives. It is the purpose of measurement in education to furnish instruments for measuring more precisely the outcomes of education, to the end that the evaluation of them may not be dependent upon insufficient and uncertain evidence.

It is, of course, necessary that the objectives of education be clearly defined or else the measuring instruments cannot be constructed. The attainment of complete clarity in objectives has been complicated by changes and additions to them introduced from time to time. Today there is much greater emphasis upon the total personality than heretofore. This means the introduction of many new objectives. At the present time we hear much about the well-adjusted emotional life, the formation of wholesome attitudes, appreciations of the beautiful, the development of interests, and the over-all picture of moral character. As soon as the objectives are clearly defined in terms of children's habits, ideals, and other behavior manifestations, measurement becomes possible. At the present time, for example, there are well-constructed inventories of emotional balance, attitude scales, tests of art and music, interest blanks, and procedures for measuring cheating, lying, and stealing.

DIFFICULTIES IN MEASURING MENTAL TRAITS

At first the difficulties of measuring the mental traits of human beings seemed insurmountable. There was such a sharp contrast between the complexity, let us say, of silent reading and the simplicity of linear distance. Even general merit in handwriting, with its elements of slant, letter formation, quality of line, spacing, and alignment, seemed complex indeed. And yet after much experiment with questions and answers in silent reading, for example, there have been secured tests which bring out the delicate shades of meaning inherent in the paragraph. If a child, then, can answer these questions (which are based on the selections read) he has achieved the objective sought in reading instruction. Handwriting, too, has yielded somewhat to a measurement of its general merit by means of a scale made up of samples of handwriting whose quality increases by steps declared equal by expert judges.

A second difficulty in measuring human traits was that of variability of the individual measured. Measurers even in the physical sciences had shown *slight* variations. Small differences, for example, in the length of an iron bar were caused by changes in temperature, and variations in the speed of sound were caused by changes in atmospheric

conditions, but these seemed trivial compared with the variations between "usual" and "best" in a child's handwriting or in the speed of reading a paragraph from one time to the next. It was discovered, for example, that far less variation in performance took place if the subjects could be induced to put forth their best efforts. Small distractions, too, were eliminated, and great care exercised in giving the same setting to a problem on subsequent occasions so that the variations from one test to another have been reduced to a known minimum.

The third problem of determining the zero of measurement, which Thorndike raised in his treatment of the fundamentals of measurement, has not been solved but has been by-passed. Mental age uses birth as the point of reference, so that a mental age of 2 years would indicate the average intellectual performance of children 2 years from their natal day. Other points of reference have been the mean of a standard group such as of all 12-year-olds. If the point of reference is clearly defined and well understood by all, the zero, or "just not any," of a trait is not of such great importance. We must remember that thermometers use both 32 degrees below freezing (Fahrenheit) and freezing (centigrade) as reference points, each of which is called zero and both of which are arbitrarily taken.

Not all difficulties of measuring human responses have been as well resolved as have the three just mentioned. *The problem of securing validity stands out at present above all others.* Validity refers to the degree of effectiveness a measuring instrument achieves in doing that which it claims or purports to do. These difficulties in securing integrity in the instrument concerned appear in achievement tests, intelligence tests, and personality inventories.

In the area of *achievement tests* the question is pretty largely one of sampling. If the habits desired, let us say, in reading are clearly defined, then a test samples judiciously the entire area. But it is easily perceivable that this procedure might omit several areas whose understanding would be highly desirable. In *intelligence testing* there is no agreed-upon criterion against which the test may be projected. If we use teachers' estimates, then the test is better than the criterion. If we use teachers' marks, we are using a criterion greatly influenced by daily attendance and personality traits. In spite of the expenditure of much energy and effort, this problem of the validity of intelligence tests remains partially unsolved. In much worse plight in regard to validity are the *personality inventories*. Let us take that of the neurotic inventory. In such an instrument are usually gathered a hundred or so items which are generally regarded as symptoms of emotional maladjustment. "Do you daydream frequently? Do you feel miserable most

of the time? Do you have spells of dizziness?" are samples. If the emotionally maladjusted always daydreamed frequently and the well-adjusted never; if the neurotic always feel miserable most of the time and the normal never; or if only the emotionally upset always had spells of dizziness and the normal never the validation process would be a comparatively simple one. But such is not the case. Perfectly normal subjects may have now and then any of the symptoms mentioned above. The validity of neurotic inventories remains an unsolved problem in the area of measurement.

Another fundamental difficulty in the area of mental measurement is that of *developing a unit of measurement which does not vary* from one situation to another. If such constant units were developed they could be added, subtracted, multiplied, and divided with no substantial errors. Three of the many attempts to secure constant units will be discussed.

In the first place, Thorndike's handwriting scale, first published in 1909, was called scientific because he apparently had discovered a unit which was the same on all occasions. To Thorndike a unit was a difference between two samples of handwriting which 75 per cent of handwriting experts had perceived. Thorndike adopted the Cattell-Fullerton theorem that differences equally often noticed are equal except when they are always noticed or never noticed. By applying this theorem to samples where the judgment of difference was never unanimous he was able to get around the last part of the theorem. Let us take as an illustration five samples of handwriting- A, B, C, D, and E. Suppose now that these samples were selected from many others because 75 per cent of the judges said that B has a higher general merit than A; 75 per cent said that C has a higher general merit than B; 75 per cent said that D has a higher general merit than C; etc. Then the differences between the samples are equal. They are equal because they are equally often noticed. In short $B-A = D-C$ or $C-B = E-D$. But 75 per cent is 25 per cent above the mean, and the statistical term which includes 25 per cent of the judgments above the mean is the probable error. The probable error was thus used as a unit of measure. The principal difficulty with this whole procedure is that the truth of the theorem on which the method is based has never been firmly established.

A second unit of measure very frequently used is the mental year, which is simply the difference between two consecutive mental ages. Mental age, first given a scientific connotation by Alfred Binet in 1908 in connection with the measurement of intelligence, has come into wide use because its meaning is so clear. But the unit "mental year" is less constant than the above-mentioned probable error. It has been demonstrated that the amount of mental growth varies from one year to the

next. In general, the unit is large during the earlier years and becomes progressively smaller from the years 12 to 20. For example, any good intelligence test will distinguish easily between the average 4-year-old and the average 5-year-old but only our most refined tests indicate a clear difference between the average 12-year-old and the average 13-year-old. It would seem therefore that the unit "mental year" varies in length from one year to the next.

A third unit of measurement which is probably more constant than the two just described is the *standard score*. McCall, who used this unit on the Thorndike-McCall reading test, called it the *T-score*. In constructing this reading test McCall struck upon the idea of using the mean of 12-year-olds as a point of reference. (It is a well-known fact that measures of any unselected group have a tendency to pile up in the proximity of the mean and to appear less and less frequently as the distance from the mean increases. This arrangement of scores is called the *normal curve*.) To obtain a standard score McCall subtracted a score from the mean and divided it by the standard deviation of the 12-year-olds. This gave a standard-deviation score. Negative scores were avoided by assuming a mean of 50. He then measured five standard-deviation units along the base line and in both directions from the mean. In this manner he had available 10 units along the base line. McCall then divided each of the 10 units into 10 smaller units. There were thus 100 units, each unit as nearly as possible equal to each other unit.

The use of these equal units can be realized when we understand that a child who increases his score from 40 to 50 T-score units has made the same gain as has another child whose score increases from 80 to 90. These standard scores have been widely used and will be discussed further on a later page.

RESULTS OF DEVELOPING UNITS OF MEASUREMENT

Granted that objectives of education have been clearly defined in terms of student reaction, and instruments which employ adequate units of measurement constructed, then there are a large variety of problems which may be attacked. Among these, method stands out prominently. For example, does the reading of a large quantity of interesting material develop a greater capacity for reading for understanding than would the more intense studying of a narrower field? Two groups equivalent to begin with in reading capacity, as based on our well-established measuring instrument, are subjected to radically different procedures under the same teacher. What is the differential effect upon the two groups of these two methods? The answer is straightforward and understandable. That method is better which

has brought about the greater change on our measuring instrument. If a large enough sample were secured to make the findings statistically reliable, the judgment could then be made that one or the other method was definitely superior for improving the understanding of reading by children at the level studied. Mind you, the judgment would not have been a valid one had not the objective been clearly defined and the measuring instrument validated on the basis of agreement with the objective. It is not difficult to see that valid judgments could be made as to the efficacy of the size of class, length of the recitation, number of books in the library, and the preparation of teachers if the trouble were taken to measure each one by means of its degree of attainment of the described objective.

Let us now suppose that in all areas of education objectives were clearly defined, and adequate measuring instruments for these objectives had been constructed, so that degrees of attainment of the objective would be immediately reflected upon the measuring instrument. Under these conditions guesswork would disappear from education. Teachers would be forced to state in terms of pupil reaction what were the objectives of each unit of work. These objectives might then be referred to a competent committee who could modify them until they were satisfactory. A committee now goes to work to construct an instrument which would faithfully reflect these objectives. The teacher and pupil would find in this instrument great benefits. The teacher could see immediately the results of her instruction. The pupil would have an incentive unsurpassed. His mark now instead of reflecting his activities in a half dozen different areas would indicate simply the degree of success attained in a single area. And while he might not be compelled to continue until he had reached an adequate score on the defined objective, he would at least *know* where he stood.

A hypothetical situation has been pictured here which exists in only a few areas of human learning and human development. It is the purpose of this book to describe objectives and instruments for measuring them. In some cases tests have been constructed with too little attention to objectives. Sometimes the objectives have been warped to fit the instrument. In many cases the objectives and instruments have not aimed at the same thing. The idea, however, cannot be condemned because of the imperfections discovered in the details of its execution.

Fairly considered and applied, this procedure will help lead us out of the area of guesswork in education. Progress comes in every area where units of work are clearly defined. In the past, in the present, and in the future, improvement in the educative process takes place most effectively in areas where objectives have been most clearly defined and measuring instruments most carefully constructed.

MEASUREMENT IN GUIDANCE

The area which illustrates the uses of valid measures in some areas and their lack in others is that of educational guidance.

The attainment of an individual on a test or examination indicates both what he has done and what he will do. If he has succeeded in a given time in learning the fundamentals of arithmetic the chances are that he will continue to learn that subject at about the same rate. Evidently, the score on a good test is indicative of present achievement and of future possibilities. For this reason test scores are very useful in guidance. Of course, the judgment made about the future progress of an individual from the available evidence, cannot be as accurate as that one made about the past. And yet, all guidance depends upon the accuracy of prediction of human behavior. The more complete the record has been up to the present, the better the prediction and the better the guidance. For best guidance the total individual must be represented. In the past, accumulated records have contained school marks in various subjects, scores on reading, intelligence tests, and a few other things. They, for the most part, have omitted records of interests, attitudes, habits of work, emotional level, adjustment to peers and teachers, etc. It can be clearly seen that many desirable objectives are not too clearly defined in the minds of the teachers, nor are there tests or measures on which they can be accurately recorded. Motives, drives, attitudes greatly influence the success or failure of individuals. No real guidance can be administered without attention to these more intangible traits. Nor can we be satisfied until both objectives and measures are well developed in these areas.

Guidance then is dependent upon the records of significant events in an individual's life up to the present time. Anecdotal records are sometimes useful because they show the whole individual in action. But the more precise measures can be made and kept, and the more all-inclusive individual records are, the better can the guidance be.

MEASUREMENT IN EDUCATION

Well-constructed, standardized measurements exist today in three large areas: (1) achievement tests, (2) intelligence tests, and (3) personality inventories and rating scales.

ACHIEVEMENT TESTS

Achievement tests are essentially improved types of examination or tests which cover an area of learning. Improvement over usual examinations and tests consists of (1) more careful selection of representative items, (2) greater care in item construction, (3) a preliminary tryout of the items selected, (4) the establishment of norms, and (5) greater

accuracy in grading or scoring. Greatest success in constructing achievement tests has come about when (1) the objectives have been clearly defined, (2) situations have been arranged so that the objectives are clearly reflected, and (3) the amounts or degrees of the objectives have been indicated in the score obtained.

Achievement tests may be divided into informal and formal. The informal tests, which are far more frequently used than the formal ones, are constructed by the teacher. Two types of them have been most common: (1) the essay test, and (2) the short-answer test. Competent teachers have been able to improve greatly both these types.

The formal or standardized tests are more carefully constructed than the informal. Their items are subject to a number of revisions and are submitted to several persons who judge their value. The selection of items which are common to textbooks or courses of study implies that a thoroughgoing canvass of materials and objectives has already been made. After all this preliminary work has been done the test in its final form is given to a large number of unselected subjects whose scores are used to establish the norms and to compute the reliability. Good constructors of achievement tests publish enough of the construction procedures so that competent judges can be certain about the test's adequacy.

INTELLIGENCE TESTS

Intelligence tests attempt to measure capacities for learning, thinking, reasoning, and so on, without regard to the materials involved. They would measure *general intelligence*. Intelligence tests may be divided, on the basis of their use, into (1) individual tests, which examine one subject at each sitting, and (2) group tests, which can be applied to many subjects at one sitting.

There are many types of individual tests, though the Binet revisions are most frequently used at present. Binet's tests, introduced into the United States in 1911 by Dr. Henry Goddard, have had many revisions and adaptations to American conditions. All these revisions use the mental age as the unit of attainment and divide it by the chronological age to compute the I.Q. Another type of intelligence test has made its appearance in recent years: the Wechsler-Bellevue. This test, intended for adult subjects and those above the age of ten, does not use the mental age but keeps the I.Q. though slightly altered in meaning.

Group intelligence tests originated from the dire need to test large numbers of army conscripts in 1917. These original tests, objectively scored, sampled much of the same behavior tested by the individual test. So many group tests of intelligence have been constructed that today satisfactory ones are available from 5 years of age to adulthood.

PERSONALITY INVENTORIES AND RATING SCALES

In this category are included attempts to measure many dimensions of personality. Self-confidence, dominance, introversion, self-sufficiency, neuroticism are samples. Most of these attempts are based on inventories in the form of questionnaires whose questions are usually answered with "Yes," "No," and sometimes with a "?." The first of these inventories was developed by Woodworth during the First World War. It consisted of 116 descriptions of mental symptoms which were to be answered "Yes" or "No." "Have you ever had fits of dizziness?" "Do you have a great fear of fire?" "Can you stand the sight of blood?" are samples of the questions used. Many other inventories with some modifications have developed from this pioneer attempt. The California Test of Personality, the Bernreuter Personality Inventory, the Bell Adjustment Inventory, and many others have been standardized.

Many behavior traits are as yet not included in standardized inventories. To get some indication of the presence of these traits in children, ratings are necessary. In such a set of rating scales as is contained in Behavior Rating Schedules¹ the scales are usually constructed of five divisions, each of which is described verbally. For example, the twenty-eighth item asks, "Is he sympathetic?" which is to be rated on the following scale:

| | | | | |
|-------------|---------------|--------------|--------------|--------------|
| | | | | |
| Inimical | Unsympathetic | Ordinarily | Sympathetic | Very |
| Aggravating | Disobliging | friendly and | Warm hearted | affectionate |
| Cruel | Cold | cordial | | |

The most recent attempts to get at the inner life of subjects in a qualitative way are the projective techniques. By presenting materials whose meaning is not too clear (unstructured), it is hoped that somehow the subject will unfold his inner life and help the observer to understand the very nature of his being. The Rorschach inkblots and Murray's Test of Thematic Apperception are good examples.

Other personality areas are those of interest, attitude, and moral character. Interest blanks may be thought of as attempts to discover those areas of interest which are directly related to success in certain occupations. Attitude scales consist of a series of statements varying all the way from complete belief to complete disbelief in some institution, idea, or race. On the church scale one can thus check a statement that the church is the noblest of our institutions or the most to be abominated. Tests of cheating, lying, and stealing are samples of attempts to know more precisely the outcome of moral instruction.

¹ Haggerty, Olson, and Wickman, *Behavior Rating Schedules*. Yonkers, N.Y.: World Book Company, 1930. Item by permission.

SUMMARY

In order to evaluate the outcomes of education, measurement is essential. It works best when objectives are clearly defined and are understood by both the teacher and the learner. Under these conditions graded situations can be arranged so that the extent of achievement of the objective can be registered upon them. Measurement is usually the introduction of a defined unit into the total. Measurements are useful for supplying facts on which better guidance may be based.

Fundamental difficulties have arisen in connection with the measurement of mental traits. The variability of human subjects, the complexity of the function measured, as well as the establishment of agreed-upon zero have proved to be difficult to solve indeed. Along with these difficulties the proof of the validity of tests, especially in the area of personality inventories, remains one of measurement's unsolved problems.

Measurement in all areas of science has been advanced by the discovery and rigid definition of suitable units which remained the same at all times. Mental age, equal-appearing units, and T-scores were cited as samples of attempts in this direction. None of these units satisfied completely the strict scientific canon of constancy. Perhaps the T-score or standard score comes the nearest to meeting this requirement. Areas in which measurements have been constructed are (1) achievement tests, (2) intelligence tests, and (3) personality inventories, which include neurotic conditions, ascendance-submission, interests, attitudes and other dimensions of personality.

QUESTIONS AND EXERCISES

1. What are the three major divisions of the process of education?

2. Why, do you suppose, was the measurement of the outcomes of education neglected?

3. Just how are objectives and measurement related?

4. Distinguish between measurement and appraisal.

5. Describe the fundamental difficulties of educational measurement. What steps have been taken to overcome these difficulties?

6. Secure an Ayres or Thorndike handwriting scale and study critically the differences in samples on each scale.

7. Why does validity receive such a

prominent place in measurement? Why is it so difficult to achieve in intelligence and personality tests?

8. Explain the difficulties in *constructing* units of measurement. What is the standard score? How is it derived?

9. How can measurement be used in guidance?

10. Describe some problems in education that might be attacked did we have satisfactory measuring instruments.

11. Describe the three large areas in which measurement has been attempted. Name one test in each area.

12. Why should measurement be made in education?

BIBLIOGRAPHY

CRONBACH, LEE J.: *Essentials of Psychological Testing*. New York: Harper & Brothers, 1949.

GOODENOUGH, FLORENCE L.: *Mental Testing*. New York: Rinehart & Company, Inc., 1949.

GREENE, EDWARD B.: *Measurements of Human Behavior*. New York: The Odyssey Press, Inc., 1941.

GREENE, HARRY A., ALBERT N. JORGENSEN, and J. RAYMOND GERBERICH: *Measurement and Evaluation in the Elementary School*. New York: Longmans, Green & Co., Inc., 1942.

———: *Measurement and Evaluation in the Secondary School*. New York: Longmans, Green & Co., Inc., 1943.

LINDQUIST, E. F. (ed.): *Educational Measurement*. Washington, D.C.: American Council on Education, 1951.

REMMERS, H. H., and N. L. GAGE: *Educational Measurement and Evaluation*. New York: Harper & Brothers, 1943.

ROSS, C. C.: *Measurement in Today's Schools*, 2d ed. New York: Prentice-Hall, Inc., 1947.

SMITH, EUGENE R., RALPH W. TYLER, et al.: *Appraising and Recording Student Progress*. New York: Harper & Brothers, 1942.

SUPER, DONALD E.: *Appraising Vocational Fitness*. New York: Harper & Brothers, 1949.

CHAPTER 2

Characteristics of Measuring Instruments

All good measuring instruments have certain characteristics in common. These characteristics have been so well developed that they may be applied as criteria of effectiveness to any old or new measuring instrument. In the area of measurement of achievement the tests of the simpler, more observable outcomes of education were the first to possess these qualities which later were found to be characteristic of all good measuring instruments. For example, Courtis's tests in arithmetic, which consisted of addition, subtraction, multiplication, and division, were observed to give nearly the same results on successive occasions and to include many of the processes involved in the four fundamental operations in arithmetic. They had therefore both reliability and validity. These same characteristics of reliability and validity were shown to apply when the outcomes of education became more complicated. The measurements of composition, silent reading, and arithmetic problems were seen to be more effective when they possessed reliability and validity. Even in the most complicated measures of ability to reason, of attitudes, of interests, and of good adjustment, progress came when they conformed to these principles of reliability and validity.

From all these attempts at measurement certain characteristics have emerged which may be regarded as being of the highest importance.

The leading characteristics of all good measuring instruments are:

1. Validity
2. Reliability
3. Administrability
4. Interpretation and comparability
5. Economy

Placed first in the list and in every way of first importance is validity.

VALIDITY

The most important question to ask about a test which is being considered for use is: "Is it valid?" When is a test valid? *What is meant by validity?* Probably a better question would be: "For what is this test valid?" If a test indicates a known amount of progress toward

an objective it is valid for that purpose. In the Courtis Research Tests in Arithmetic, Addition consists of adding sets of nine three-place numbers. The score is in terms of speed and accuracy. This test, then, is valid for measuring speed and accuracy in column addition. It is not valid for measuring the addition of fractions or decimals or denominate numbers. It is valid for a particular purpose. Some have said, "A test is valid in proportion as it measures what it purports to measure." One author emphasizes our knowledge of what a test measures as being an indispensable characteristic of validity: "A test is valid, to the degree that we know what it measures or predicts."¹ There is a logical fallacy here, since we might *know positively* that a test does not satisfy its claims. It might be truer to say that *a test is valid in proportion as it measures well what is desired to be measured*. The phrase "measures well" implies an empirical trial of the test with an adequate sample of subjects and computations to indicate the degree of success it had achieved in measuring the desired outcome. If, then, the instrument which is chosen reflects accurately the degree of attainment of a defined objective it is valid for that purpose. To ensure this validity careful test builders exert great care (1) in the construction of the test, and (2) in correlating it with some external criterion. We might call the first of these *internal validity*; the second, *external validity*.

INTERNAL VALIDITY

Achievement Tests

Internal validity refers to the care with which the items of the test are selected and arranged. The elements which make up a test are constructed after a consideration of the agreed upon objectives. The items are carefully written, judged by a jury of experts, and then tried out upon a small sample of subjects. Ambiguities and misunderstandings are sure to appear in connection with certain items. These items are modified in statement or omitted entirely. Sometimes, even at this late date, further revisions are made before the test assumes its final form.

If our objective were to make the most valid test for an elementary algebra class, the teacher would be the best one to do it. He would know exactly the areas he had taught, the objectives he had in mind. He might analyze the areas into the processes employed and then construct a test which contained samples of all the algebraic processes, with each process being represented at three or four different levels of difficulty. If such a test were carefully constructed it would reflect accurately progress in the mastery of the algebraic processes studied

¹ Cronbach, Lee J., *Essentials of Physiological Testing*, p. 48. New York: Harper & Brothers, 1949.

and the defined objectives. In such a test the *curricular* or *internal validity* would be satisfactory. For obtaining the curricular validity for this particular subject, this procedure has no rival.

Frequency of Occurrence

In contrast to the teacher's test of specific subject matter a *standard test* over the same area would base its items on *subject matter common to courses of study and popular textbooks*. The procedures used in constructing such a test indicate the method. Let us see how it worked in one case. In constructing the literature section of the Unit of Attainment Test, the literary samples were selected from lists recommended by state courses of study. A list of the better state courses of study was made for the author by one member of our department, M. R. Trabue,¹ who had at that time been investigating state courses of study. This list was then rated by two other competent persons. With this list of 10 courses of study in hand, prose and poetry selections were made which were common to at least 9 out of the 10. Multiple-choice items for each selection were then constructed.

In the construction of other tests, many devices have been used to find the most frequently used materials. In one case, a pool of items was made from those common to a list of textbooks regularly used in that area. In another, questions from a sequence of examinations have been inspected. In a series of examination questions, some questions in slightly different form occur more than once. These have been used as bases for test construction. The use of frequency of occurrence in textbooks or courses of study as a criterion tends to neglect local materials introduced for interest and to perpetuate common facts in the test. Unless the new material introduced into one textbook were incorporated generally into others, it could not appear in the test. At times, the undue influence of test items on teachers has tended to discourage not only experimentation with the curriculum but also the introduction of materials gathered from the locality.

This implied tendency of a certain type of standardized test to "freeze" the content of the curriculum may be largely avoided by constructing a test of the more permanent aspects of education. Thus the test may not be naïvely concerned with the mere reproductions of facts but may deal with the interpretation of these facts embedded in a new situation. In science, for example, this would mean tests of understanding scientific method such as the formulation and testing of hypotheses or the solution of problems unlike any that had been studied. With this type of item the criticism that objective tests encourage memorizing specific facts disappears.

¹ Now of Pennsylvania State College.

Judgment of Experienced Observers

The use of the judgment of experienced observers as a criterion against which to measure a test's validity is nicely illustrated in the construction of the Iowa Silent Reading Tests. This was directly influenced by *A Survey of a Course of Study in Reading*.¹ This investigation listed and analyzed the characteristics which are ordinarily met in typical reading situations (Table 1).

TABLE 1. READING ABILITIES VS. READING TEST
Horn and McBroom's List of
Reading Abilities

| Reading Abilities | Iowa Silent Reading Test* |
|---|--|
| 1. Skill in recognizing new words | Test 1. Word meaning Part A. Social science Part B. Science Part C. Mathematics Part D. English |
| 2. Ability to locate material quickly. Involved use of index, table of contents, dictionary, card files, etc. | Test 2. Location of information Part A. Use of the index Part B. Selection of key words |
| 3. Ability to comprehend quickly what is read | Test 3. Paragraph meaning Part A. Science Part B. Poetry Part C. Political science |
| 4. Ability to select and evaluate material needed | Test 4. Paragraph organization Part A. Selection of central idea Part B. Outlining |
| 5. Ability to organize what is read. Involved summarizing, ordering of topics, discovery of related material, and outlining | Test 5. Sentence meaning. (Set of sentences of increasing difficulty to be answered by "Yes or No") |
| 6. Remembrance of material read | Test 6. Rate of selected reading Part A. In reading science material Part B. In reading political science material |
| 7. Knowledge of sources | |
| 8. Attitude of attacking reading with vigor | |
| 9. Attitude of proper care of books | |

* Test numbers changed.

If one compares the two columns of the table, one sees that while the test follows this analysis of reading pretty closely, it emphasizes

¹ Horn, Ernest, and Maude McBroom, *A Survey of a Course of Study in Reading*, Extension Bulletin No. 93, College of Education Series No. 3, University of Iowa, 1924.

more the facility in reading with comprehension than the many other uses to which reading is put. Thus the test emphasizes "the ability to comprehend quickly what is read" in two selections—one from science and one from literature—as well as in the understanding of sentences. Word knowledge is well represented as well as the looking up of items in indexes and the speed of reading. Attitudes, knowledge of sources, and the proper care of books are omitted. We can see that this excellent test of reading is not completely valid. Such a procedure for selecting items is less static than the preceding and includes some aspects of social utility.

Social Utility

It is inconceivable that the criteria thus far presented for selecting items for a test should have been entirely devoid of social utility. Even if the criteria used implied the presence of social utility its significance must not stop at mere implication. Social utility is then used here as a separate criterion although it is related to all the others. Here is a test in spelling, for example, which selects its words because of their frequency in private correspondence, or another, because of the frequency of references in reading, or yet another, because of the number of times words are misspelled. At least one test in home mechanics has been based on a course of study which was composed of activities engaged in while mending the things around the home.

But until educational objectives are dominated by the ideal of social utility in their formulation, the measurer is helpless. Remember that a good measuring instrument is valid only in so far as it indicates the degree to which an agreed-upon objective has been reached.

Psychological and Logical Analysis

One of the best illustrations of a slightly different emphasis upon validating criteria appears in the report of the Evaluation Staff of the Progressive Education Association.¹ Their procedures illustrate precisely what is meant by psychological analyses in test construction. In this investigation clear-cut objectives were first decided upon. The 30 participating schools entered into this project by setting forth the objectives of education which their respective staffs had worked out. This rather long list was studied by the Evaluation Staff and consolidated into ten objectives, as follows:

1. Methods of thinking
2. Useful study skills and work habits
3. Social attitudes

¹ Smith, Eugene R., Ralph W. Tyler, *et al.*, *Appraising and Recording Student Progress*. New York: Harper & Brothers, 1942.

4. Wide range of significant interests
5. Increased appreciation of music, art, and literature
6. Social sensitivity
7. Better personal and social adjustments
8. Acquisition of important information
9. Physical health
10. Consistent philosophy of life

After the objectives were agreed upon, the search began for types of materials through which these objectives are expressed. A method of scoring the reactions was then worked out so that a more precise agreement between the defined objective and the evaluating instrument could be realized. In the final step, a careful interpretation of the whole procedure was developed. In the book just referred to, these three steps—(1) finding materials, (2) discovering means of registering accurately the reactions of subjects, and (3) checking the objective against the results thus achieved—were employed for each of the 10 objectives listed above. Here we will summarize only the procedure used in **evaluating methods of thinking.**

In the first instance, "methods of thinking" was defined more clearly. It was agreed that methods of thinking included at least four abilities: (1) ability to interpret data, (2) ability to apply principles of science, (3) ability to understand the nature of proof, and (4) ability to formulate hypotheses. The ability to interpret data involves (a) ability to perceive relationships in data, and (b) ability to recognize the limitations of data. In this manner each of the four abilities was analyzed into smaller, more understandable parts which could be clearly perceived and whose expression could be observed in selected materials. In validating and appraising such an objective as methods of thinking there were no limits to the types of material that could be used. The form, however, must be *new* to the subject, or else his act would be a simple one of memory. The social sciences and natural sciences offered satisfactory material for this purpose, and so selections were made from them. An illustration of the procedure is provided by a sample exercise (Problem 1) from Form 2.52:¹

These data alone

- (1) **are sufficient to make the statement true.**
- (2) are sufficient to indicate that the statement is probably true
- (3) are not sufficient to indicate whether there is any degree of truth or falsity in the statement.
- (4) are sufficient to indicate that the statement is probably false.
- (5) are sufficient to make the statement false

¹ Smith, E. R., R. W. Tyler, et al., *Appraising and Recording Student Progress*, pp. 52-53. New York: Harper & Brothers, 1942. Quoted by permission.

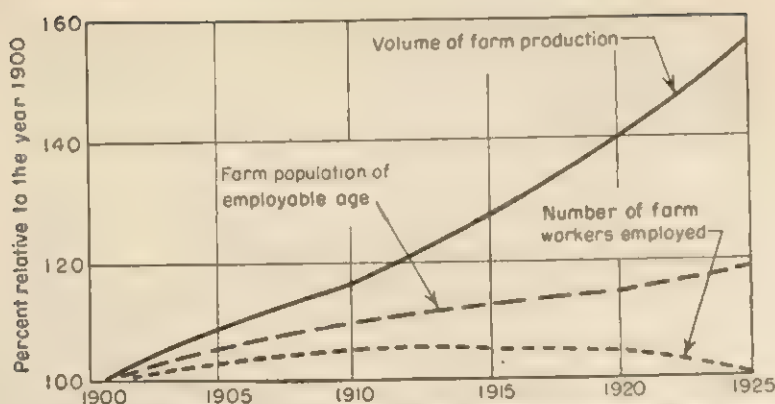


FIG. 1. Problem 1. This chart shows production, population, and employment on farms of the United States for each fifth year between 1900 and 1925.

Statements

1. The ratio of agricultural production to the number of farm workers increased every five years between 1900 and 1925.
2. The increase in agricultural production between 1910 and 1925 was due to more widespread use of farm machinery.
3. The average number of farm workers employed during the period 1920 to 1925 was higher than during the period 1915 to 1920.
4. The government should give relief to farm workers who are unemployed.
5. Between 1900 and 1925, the amount of fruit produced on farms in the United States increased about fifty per cent.
6. During the entire period between 1905 and 1925 there was an excess of farm population of employable age over the number of people needed to operate farms.
7. Wages paid farm workers in 1925 were low because there were more laborers than could be employed.
8. More workers were employed on farms in 1925 than in 1900.
9. Since 1900, there has been an increase in production per worker in manufacturing similar to the increase in agriculture.
10. Between 1900 and 1925, the volume of farm production increased over fifty per cent.
11. Farmers increased production after 1910 in order to take advantage of rapidly rising prices.
12. The average amount of farm production was higher in the period 1925 to 1930 than in the period 1920 to 1925.
13. Between 1900 and 1925 there was an increase in the farm population of employable age in the Middle West, the largest farming area in the United States.
14. Farm population of employable age was lower in 1930 than in 1900.
15. The production of wheat, the largest agricultural crop in the United States, was as great in 1915 as in 1925.

From such a test we may secure eight different scores: (1) general accuracy, (2) probably true or probably false, (3) insufficient data, (4) true-false, (5) omitted, (6) caution, (7) beyond data, and (8) crude errors. Items 1, 2, 3, 5, 7, and 8 are self-explanatory. Item 4, true or false, gives the percentage of times the subject recognized a true statement as true and a false statement as false. Item 6, caution, refers to the withholding of the degree of truth which the makers of the tests would allow. In thus producing an analyzed score the test could focus the teacher's thought on the weak and strong points in the student's ability to think.

If one of the objectives striven for by teachers in instructing high school students is the ability to apply principles of science, and if this objective is analyzed and areas discovered where the application is feasible, then the degree to which the objective has been reached may be measured. The teacher can then decide whether or not his teaching procedures have been effective for this purpose, and the student can be properly guided into activities which demand the amount of scientific generalization achieved by him. This procedure in test construction is interesting because the whole process from objective to the evaluation of the instrument is set before us. Moreover, the attempt was made to develop instruments in areas where no satisfactory instruments already existed. Finally, it is instructive to those of us now working on validity because the authors really set down validity as the first and foremost of their criteria in the construction of their tests.

Intelligence Tests

In constructing intelligence tests there is no common pool of information from which questions can be drawn. Items, in general, are selected because they are drawn from the common environment, because they are passed by an increasing number of subjects with increasing age, or because an increasing percentage is passed as I.Q.s increase from 90 to 100 to 110. For example, in the Stanford Revision of the Binet-Simon tests, if a smaller percentage of children whose I.Q.s were 110 passed the item than of those with I.Q.s of 90, the item would not be selected. Items used in the Terman-Merrill Revision were also correlated with the test as a whole. If the new item did not agree well with a score based on the total of items, it was eliminated. A different way of selecting items appears in the work of Maurer.¹ For a long time it had been known that tests given in the early years of life did not well predict standing in the later years. Maurer was able to study the predictive

¹ Maurer, Katherine M., *Intellectual Status at Maturity as a Criterion for Selecting Items in Preschool Tests*. Minneapolis: University of Minnesota Press, 1946.

capacity of items of the Minnesota preschool tests by correlating them with a group intelligence test given in late adolescence. She demonstrated that tests could be selected which would predict later standing on group tests of intelligence. Thus a new procedure for selecting items was developed.

Aptitude Tests

Items for aptitude tests have been selected by a psychological analysis of the factors involved, as in Seashore's Measures of Musical Talent, or by a correlation of each item with some criterion of success. The latter procedure, most generally used at present, appraises what is known as *external validity*. If we were selecting items for a clerical-aptitude test, internal validity would demand that each item correlate well with the total score. Suppose we should take the highest 27 per cent and the lowest 27 per cent from scores made on our Total Test. Any item that was passed by a much larger percentage by members of the highest group than by those of the lowest group would be a suitable one. If an item were passed by a larger percentage of subjects in the lower group than in the upper it would not be discriminative and therefore could not be used.

EXTERNAL VALIDITY

However well a test is prepared there is no certainty of its usefulness until it is tried out by comparing it (1) with actual achievement in a practical situation, or (2) with other measures of the same area. After all, there is usually some measure outside the test itself against which this measuring instrument may be projected. These outside measures are called *criteria*. If satisfactory criteria could be established for all tests, their validity might be efficiently appraised.

Achievement Tests

The criteria against which we attempt to measure achievement tests are usually much less effective measures of achievement than the tests themselves. One might use teachers' marks as criteria of success, but they are compounded of many elements in addition to achievement in school subjects. Teachers' ratings of achievement in reading or arithmetic make the criterion purer but add to the problem the unreliability of rating. Achievement tests have rather high correlations (.70 to .80) with intelligence-test scores, but these have many other components than achievement. For this reason, constructors of achievement tests are depending more and more on *curricular validity*.

Intelligence Tests

It is also difficult to discover adequate criteria against which to measure intelligence tests. One criterion which has sometimes been

used is the average rating of three or four competent persons. The intelligence of some hundred children is rated by three persons who know them well. The average of these ratings is computed, and with this average the scores of the test are correlated. Another criterion with which group test scores have been compared is the individual test. For example, a group test of intelligence may be correlated with an individual test which has been long established, such as the Stanford-Binet. On one occasion, the author correlated the scores of four group tests of intelligence with the Stanford Revision of the Binet-Simon tests, thinking that perhaps the one with the highest correlation with this individual test might be a more efficient measuring instrument for intelligence.¹ School marks, in spite of the multiplicity of factors which sometimes enter into their composition, have been used as criteria both for achievement tests and for intelligence tests. In one case (Terman, 1916) the coefficient of .48 was given as existing between the Stanford Revision intelligence test and school marks. In general, the correlation between average school marks and intelligence-test scores would range from .40 to .60.

An illustration of validation through statistical procedures may now be presented. The author¹ had in mind the determination of the highest validity among four group tests of intelligence—Army Alpha, Terman Group, Otis Advanced, and Miller. In this study each of the four tests was measured against four important criteria: (1) Stanford-Binet, (2) teachers' ratings of intelligence, (3) school marks, and (4) a composite made up of a combination of all four group tests. Each of the 64 students was tested with all four group tests as well as with the Stanford Revision of the Binet test. The teachers' ratings of intelligence represented the average ratings of four critic teachers who knew the pupils well. The school marks were averaged for each student. The comparisons in all instances were made by means of Pearson's Coefficient of correlation.

1. The coefficients of correlation computed with the scores on the group tests and the mental ages secured from the Stanford-Binet were in the neighborhood of .68 for three of the group tests and .53 for the other. These results indicate substantial or marked correlations, but in no case is the correlation a high one. As measured by this first criterion these four group tests do measure a considerable amount of ground common to the Stanford-Binet, but there is an area of unlikeness between any one group test and the individual test.

2. In the case of teachers' ratings of intelligence, the correlations

¹ Jordan, A. M., "The Validation of Intelligence Tests," *Journal of Educational Psychology* (1923) 14:348-366, 414-428.

computed with the group tests ran from .60 to .70. Here again the agreement is substantial between group tests and what competent persons judge to be the presence of intelligence.

3. When school marks were correlated with each of the four group tests, the coefficients varied around .47, with the lowest being .45 and the highest .49. According to these figures intellectual factors measured by our group tests entered into the securing of school marks to only a moderate degree. The correlations are, however, of about the same size as that found for the Stanford-Binet and school marks ($r = .48$).

4. Finally, when the group tests were correlated with a composite score made up of all of them combined, there is an entirely different size of correlations, for now they are .90 and above. This signifies that each group test is measuring about the same characteristics as their combination. The fact that each group test's score was included in the composite tended to raise the size of these coefficients.

In this same article, many of the correlation coefficients which other investigators had previously computed between each group test of intelligence and the four criteria mentioned above were collected. For example, between Army Alpha and high school marks 26 coefficients were found, and 35 with college marks. The average of these relationships between Army Alpha and school marks was .38. In this manner, when all correlation coefficients in which a group test of intelligence entered are collected, a great deal is known about its validity. Truly, *a test is known by its correlations.*

Aptitude Tests

Aptitude tests have used measures of achievement in a realistic situation as criteria to indicate the presence of external validity. A good illustration of the development of a satisfactory criterion appears in the standardization of the Minnesota Mechanical Ability Tests. The criterion which was finally utilized was the quality of mechanical work which students produced in junior high school. This quality was arrived at by direct observation and inspection of the work, by actual measurement of the product, and by judging the output by refined scales. Time has shown that the criterion was a good one and that the time consumed in constructing an adequate criterion was well spent. Throughout this text examples of criteria used will be illustrated whenever tests are discussed.

Recent Trends in Test Validation

In recent years there have been no fundamental changes in studying the validation of intelligence tests. There has, however, been extension in three directions: (1) one test is studied at a time, (2) the number of criteria against which the test is projected has been increased, and

(3) there has been great interest in the validity of individual parts of tests. At the present time the validity of an intelligence test is determined by its correlations with the following criteria:

1. School marks. Average school marks and marks in individual subjects are utilized. Sometimes scores obtained from educational achievement tests are used.

2. Other intelligence tests, especially those which have been used a long time and about which much is known.

3. Mechanical, clerical, and artistic ability as measured by tests in these fields.

4. Success on the job. There has been much interest here in connection with the use of tests in guidance. Success in salesmanship and teaching are examples.

5. Amount of education which individuals have achieved. The correlations are made with the highest grades achieved in school.

6. Length of time remaining in school or progress through school.

7. Many other miscellaneous criteria.

Against all these criteria are projected both the test as a whole and each of its major parts.

Since many of these criteria of validity are considered when the various intelligence tests are treated in this text, a few illustrations only will be given here.¹ Thus two investigators found that correlations between the A.C.E. (American Council on Education) Psychological Examination and the school marks of the University of Chicago freshmen ranged from .48 (biological sciences) to .57 (social sciences).² This same A.C.E. test correlated from .58 to .67 with the Terman-Merrill Revision.³ Another student computed a correlation of .62 between the A.C.E. and name checking and one of .26 between the A.C.E. and number checking.⁴

Vitiating Factors in Validity

The validity of a measuring instrument is sometimes reduced in effectiveness by impurities which creep either into its content or into its administration. Some of these factors are:

¹ For a much more exhaustive treatment of this topic, see Super, Donald E., *Appraising Vocational Fitness*, Chap. VI. New York: Harper & Brothers, 1949. See also Seagoe, M. V., "Prognostic Tests and Teaching Success," *Journal of Educational Research* (1945) 38:685-690.

² Shanner, W. M., and G. F. Kuder, "A Comparative Study of Freshmen Week Tests given at the University of Chicago," *Educational and Psychological Measurement* (1941) 1:85-92.

³ Manuel, H. T., et al., "The New Stanford-Binet at the College Level," *Journal of Educational Psychology* (1940) 31:705-709.

⁴ Super, Donald E., "The A.C.E. Psychological Examination and Special Abilities," *Journal of Psychology* (1940) 9:221-226.

1. In some cases a test item which seems to be a good measure of one objective, measures *another* also. An item in an intelligence test might conform to all criteria used in its selection but, because it depends on reading, would make a poor item for measuring the intelligence of slow readers.

2. In certain tests of clerical ability, speed is the dominant factor in making a good score. Some teachers have so insisted on accuracy that when their students took this test of clerical ability so dependent on speed, they could not force themselves to speed up. With this group of students the test was invalid for measuring rate.

3. In the Strong Vocational Interest Blank the subject votes L-I-D (like, indifferent, dislike) on most of the items. It was thought that subjects in the vast majority of cases would use either L or D and would use I only when they simply could not decide. Some subjects, however, are unable to make affective judgments of either L or D and use I on a very large number of items. For this group no clear direction of vocational interest can be secured from the administration of the blank.

4. Through experimenting with the true-false technique used in constructing test items it was discovered that students when in doubt mark the item "True." Such items, if true would be correctly marked. If false, they would be incorrectly marked and would therefore be a more precise measure of the subject's knowledge. In one study Cronbach¹ showed that the reliability of his "false" items was .72, that of his "true" items, .11. False items, then, were more reliable and more useful.

In short, a variety of unpredictable human factors sometimes prevent the item from measuring those processes for which it was prepared and thus invalidate it for the purpose at hand.

RELIABILITY

A good measuring instrument must of necessity possess the characteristic of reliability. Reliability implies precision or accuracy. When a test possesses high reliability its results vary little from one test to another. It gives nearly the same results on two successive occasions. Suppose that a child receives a mental age of 6 years and 10 months (6-10) on one testing of the Terman-Merrill Revision and 7-0 at the next which is given one week later. These are accurate results, and if 100 pupils were tested on two occasions a week apart and registered such small variations for each of the 100 subjects involved, the test would be designated "highly reliable." *In validity the emphasis is on a*

¹ Cronbach, Lee J., "Studies in Acquiescence as a Factor in a True-False Test," *Journal of Educational Psychology* (1942) 33:401-415.

test's agreement with the objective; in reliability, upon agreement with itself. In terms of the oft-worked analogy of linear measurement, the yardstick's validity is determined by its agreement with the standard yard in our National Bureau of Standards, its reliability, by its agreement with itself. A certain board's length remains at $16\frac{3}{4}$ inches through three successive measures, a fact which indicates the measuring instrument's lack of variation (its reliability). Further understanding of reliability may be achieved by following carefully the four methods which are used for computing it.

METHODS FOR COMPUTING RELIABILITY OF TESTS

In three of the four methods the technique used for measuring reliability is the coefficient of correlation.

1. *The repetition of the same test.* When there is only one form of a test, reliability may be measured by the correlation between the scores received from two administrations of the same test. Each of 100 subjects, say, would possess two scores received on the same test given at different times. The reliability would be obtained by computing the coefficient of correlation between them. One can readily see that when the same test is repeated some of the children will remember the items from its first administration and some curious ones will have looked up the answers or asked their parents. One question which always arises relates to the amount of time which should elapse between the two testings. If only a short time elapses, then the memory factor may be quite large; if a long time, the scores achieved are affected by the amount of growth which has taken place during this period. There is also the problem of the variable physical and emotional reactions from one test to another, since a child who is well oriented on one occasion is in a state of emotional excitement on another because an aunt has died or perhaps because Christmas is in the offing. For all these reasons this test-retest technique is now rarely used.

2. *The use of two forms of the same test.* If a test has two equivalent forms with about the same mean, the same variation, and the same selection and difficulty of items, then the correlation between these two forms constitutes one of the best methods of computing reliability. In general, subjects, because of the familiarity with the form of the questions and the similarity of content between the two forms, tend to make a slightly larger score on the second test. Since there is a tendency for all subjects to increase their scores by a small amount, the correlation would not be affected. All that was said about the changes in emotional level, attitudes, and interests in the case of the test-retest technique is also true here. One might say that the reduction of the coefficient from 1.00 is an indication of the effect of chance errors just described, since

constant errors do not affect the coefficient. Chance errors produce changes in a score's position either up or down and thus reduce the size of the reliability coefficient.

3. *The odds-even or split-half method.* This method does not involve the repetition of a test either in the same form or in a different form. In applications of this procedure, after the test is given the items are divided into equivalent parts or tests by placing the correct odd items in one part and the correct even items in the other. If the items of the test have been well scaled in difficulty in the first instance, two equivalent parts can be constructed. These two parts are now treated as two forms of the same test and the coefficient of correlation computed between them. We thus have a reliability coefficient based on a test half as long as the original one. How reliable would a test be which is just twice as long as the half-tests just now constructed? To answer this question we use the Spearman-Brown prophecy formula:¹

$$r_{nn} = \frac{nr_{11}}{1 + (n - 1)r_{11}}$$

where r_{nn} is the correlation between n forms of a test and n parallel forms and r_{11} is the reliability coefficient. In this case r_{11} will be $r_{\frac{1}{2}\frac{1}{2}}$

which is the odds-even coefficient and is assumed to be .80. The whole test, being twice as long as the half, would then have the following reliability:

$$r_{nn} = \frac{2(.80)}{1 + (2 - 1).80} = \frac{1.60}{1.80} = .89$$

The total test's reliability (r_{nn}) would thus be .89. Many students of testing prefer this procedure since it eliminates the changes in emotional level, the ill effects of memory, and the bothersome problem as to how long the period between testings should be. Garrett points out that the prophecy formula is valid only when the test items in the two parts cover the same ground, are of equal range or difficulty, have the same average scores, and are as reliable in one part as in the other. By empirical procedures it has been demonstrated that a test actually twice as long will have the same coefficient as the one predicted by the formula. It is reported that a correlation coefficient thus derived is larger than one computed from two forms but probably is the *true reliability*.

¹ Garrett, Henry E., *Statistics in Psychology and Education*, 3d ed., pp. 387-391. New York: Longmans, Green & Co., Inc., 1947.

4. *Reliability without correlation (Kuder-Richardson technique).*¹ A newer technique for computing reliability has been developed which requires only three sets of facts: (1) the number of items in the test (n), (2) the standard deviation of the test as a whole (σ_t), and (3) the arithmetic mean of the test scores (M_t). One formula frequently used in the Kuder-Richardson technique is

$$r_{tt} = \frac{n}{n-1} \frac{\sigma_t^2 - n\bar{p}\bar{q}}{\sigma_t^2}$$

$$\bar{p} = \frac{\text{arithmetic mean of test scores}}{n} = \frac{M_t}{n}$$

$$\bar{q} = 1 - \bar{p}$$

Suppose we had a test such as the Otis Advanced Intelligence Test with 212 items, whose standard deviation was 25 and mean 150. Then

$$r_{tt} = \frac{212}{211} \cdot \frac{625 - 212(.71)(.29)}{625} = .93$$

$$\bar{p} = \frac{150}{212} = .71$$

$$\bar{q} = 1 - .71 = .29$$

This formula posits several assumptions which are not always true. One of these assumptions is that all items are of the same difficulty. In so far as this is not true and there is variation of difficulty among the items, the size of r is reduced. However, Garrett points out that this formula will give a satisfactory approximation to a test's reliability even when the test items cover a wide range of difficulty.² Two other assumptions (1) that the item intercorrelations are equal, and (2) that the test items measure essentially the same ability—must be true if this formula is to give a very accurate reliability coefficient. Its results are always lower than the other methods, so that the true reliability is at least as high as the one this method gets. For these reasons, this procedure is recommended only when a rough estimate of reliability is demanded and when a quick answer is imperative. Since several factors influence the reliability, the student will observe carefully (1) the procedure used in computing the coefficient, (2) the representativeness of the population, and (3) the standard deviation of the population used.

FACTORS WHICH AFFECT RELIABILITY

What we record and measure are human reactions. These responses vary greatly from time to time even to the same situation. So much

¹ Kuder, G. F., and M. W. Richardson, "The Theory of the Estimation of Test Reliability," *Psychometrika* (1937) 2:151-160.

² Garrett, *op. cit.*, pp. 385-386.

depends on interest and effort, on physical conditions, on emotions, and on thought processes already in progress that even under the best conditions there would be some variation from one time to the next. Even if the measuring instrument were perfect and the conditions of the testing were ideal in every particular, there would still be variation in the subject's responses, a fact which would lower the reliability. Whenever reliabilities are reported for a test, it is understood that testing was done under good conditions by a person who knew children and who knew the importance of carrying out accurately the written instructions of the test.

1. *Factors which reduce reliability.* We can divide these factors into three groups. First, the *subject*—all those factors which cause variation in his reactions reduce reliability or accuracy. Here we have variations in motivation, in emotional balance, in physical level, and in thought processes already established. Second, the *tester* sometimes does not follow the instructions exactly, is careless about the time allowed for each test, does not see that the young child understands the problem before the test proceeds, and is not keenly sensitive to the possibility of cheating. Sometimes the tester, too, is a “deadpan” who somehow or other does not inspire children to want to work. What is desired on all tests is the *best* which subjects can produce. Any variation from the best is apt to lower reliability. In the third place, the *scorers* may not be accurate in their scoring. It is so easy to make mistakes in scoring. Particularly is this true when the test itself allows the scorers some discretion in interpreting the answers. On the Stanford Revision, for example, at year VII there is a diamond to be drawn which must be judged as *passed* or *failed*. In many cases the judgment is easy but often there is disagreement among equally competent observers. When words are to be written in to complete sentences, such as “_____ should prevail in libraries and churches,” a bright student will sometimes suggest a word which was not intended by the test builder and which therefore will be interpreted differently by different scorers.

2. *Factors which increase reliability.* The factors which increase reliability are, first of all, the opposite of the conditions which reduce it: good motivation which extends throughout the test, emotional calmness, careful administration, and effective scoring. Secondly, the lengthening of the test affects directly its reliability. This fact might have already been inferred from the fact that a whole test is more reliable than a half of one. Sometimes a test constructor has this problem: “My test, which has 75 items and takes 30 minutes to administer, has a reliability of .85, but I want a reliability of .95. How much longer will my test have to be to secure a reliability of .95?” Again the Spearman-Brown formula becomes useful, but now we have

to solve for n :

$$r_{nn} = \frac{nr_{11}}{1 + (n - 1)r_{11}}$$

This becomes

$$.95 = \frac{n(.85)}{1 + (n - 1).85}$$

which when solved for n gives 3.5. He would need then a test of 3.5 times 75 (or 262) items in length and one that would take 1 hour and 45 minutes to give. While it might be more efficient in this case to work on the internal consistency and structure of the test rather than merely to lengthen it, still the importance of the mere length of the test for reliability is clearly demonstrated.

3. *The range of the subjects.* This too affects the reliability. Let us take a case where only three subjects were included: one a genius, one an average child, and one an idiot. In all tests the variation of the idiot would never be so great as to exceed the average individual, nor would the average child score as high as the genius or as low as the idiot. In this case the reliability would be represented by 1.00 on the poorest of tests. Make the range great enough and your reliability is practically perfect. But such tests would not be valuable because their scores would vary too much from time to time. What we want is a test which will distinguish between subjects closely alike in, say, their intelligence. We need a test which reveals correctly the members of a single class where the variation in scores may be small. In short, reliability computed from a population composed of the members of three grades would necessarily be higher than from a population drawn from a single grade. Kelley's formula may be applied:¹

$$\frac{\sigma_s}{\sigma_L} = \frac{\sqrt{1 - r_{LL}}}{\sqrt{1 - r_{ss}}}$$

where L = large group

s = small group

If we would secure from a single sixth grade with an σ_s of 10 and a r_{ss} of .60, what would the correlation become if we used three grades with a standard deviation of 20? This becomes in the formula:

$$\frac{10}{20} = \frac{\sqrt{1 - r_{LL}}}{\sqrt{1 - .60}}$$

$$r = .90$$

¹ Kelley, Truman L., *Statistical Method*, p. 222. New York: The Macmillan Company, 1923.

From this discussion it is clear that test constructors should define meticulously the variation of the subjects from whom the reliability was computed. In general we can say that the *variability of the standardizing population should correspond to the variability of the class or grade for which the test is going to be used*. If discrimination is required for a single grade, then the reliability should be computed from a population composed of members of a single grade. In Pintner's Verbal Series of Intelligence Tests the intermediate scale's reliability is computed from children within the age range of one year. This is an excellent illustration of correct procedure.

INTERPRETATION OF RELIABILITY COEFFICIENTS

The practical questions arising out of the previous discussion are "What do these coefficients mean?" and "How large must the coefficient of reliability be to be satisfactory?" In the first place, the answer to the question depends on the accuracy required for the purpose at hand. If one wants merely to distinguish between two groups of individuals, a reliability of .50 will be satisfactory, but if he wishes to distinguish between individuals in such a way that the score indicates an accurate estimate of an *individual's* present status and some indication of his future achievement, the correlation indicating reliability must be much higher. In this latter case, the coefficient should be above .90, as much above as we can get. Some of our best achievement tests have coefficients as high as .96 or .97. The reliability of the Terman-Merrill Revision for all I.Q.s computed above the age of 6 years is .93; for the feebleminded the reliability is .98 (the highest reported). It is not at all unusual for intelligence tests or achievement tests constructed in recent years to report a coefficient as high as .95. This is true when due regard has been paid to the variation of the subjects used in the computation. In these two areas one should not choose a test whose reliability is below .90, certainly not for use in school. In the areas of interests, attitudes, neuroticism, ratings, etc., one has to be satisfied with instruments which are not quite so reliable.

From the reliability coefficients one may calculate the efficiency of one form of a test in forecasting scores on another form. The coefficient which is used to calculate this relationship has been called the *coefficient of forecasting efficiency*.

$$E = (1 - \sqrt{1 - r^2})100$$

If $r = .85$, then $E = 47$ per cent

If $r = .90$, then $E = 56$ per cent

If $r = .95$, then $E = 68$ per cent

If $r = .98$, then $E = 80$ per cent

If $r = .50$, then $E = 13$ per cent

Notice the difference in efficiency between a reliability of .90 and one of .95, an increase of 13 per cent. Probably most surprising of all is the difference in efficiency between a reliability of .95 and one of .98. This increase in reliability of .03 is accompanied by an increase in efficiency of 11 per cent.

Probably the most practical and perhaps the best interpretation of all arises out of the concept of the variation of the obtained score. After all, we want to know *how much confidence we can place in the individual score*. In brief, if a subject were tested 100 times on this test until his true score were obtained, how much is his present score likely to vary from this true score? The formula used for this calculation is:

$$\sigma_{100} = \sigma_1 \sqrt{1 - r_{11}}$$

where σ_{100} is the standard error of an obtained score, r_{11} is the reliability coefficient, and σ_1 is the average of the standard deviations of the two forms. Suppose that the score of an individual on a 100-item test is 60, the σ_1 is 7, and the coefficient of reliability is .85. Then

$$\begin{aligned}\sigma_{100} &= 7 \sqrt{1 - .85} \\ &= 7 \times .39 \\ &= 2.73\end{aligned}$$

or 3— in round numbers. We now apply this 3 to the score on the test, 60 ± 3 . This means that the chances are 68 in 100 that the true score lies between 57 and 63 and more than 99 in 100 that the true score lies between 51 and 69; *i.e.*, between 60 and three times its standard error on one side and between 60 and three times its standard error on the other.¹ Let us take an actual case from the Terman-Merrill Revision. For an I.Q. of 130 the standard error of an obtained score is 5.24, or 5 in round numbers. If we apply that to 130, then we get 130 ± 5 . The chances are 68 in 100 that the true score lies between 125 and 135 and 99.7 in 100 that the true score lies between 115 and 145. This seeming complication at first soon disappears and with practice we think to ourselves "Score 85, standard error 10—not so good," or "Score 85, standard error 3—good," because we know that the variation in the last instance is small indeed. The standard error on the Stanford Achievement Test is 2 months. We thus say Mary has an educational age of 8 years and 6 months (8-6) plus or minus 2 months. The chances are 68 in 100 that Mary's true educational age is between 8-4 and 8-8 and 99 in 100 (practical certainty) that it lies between 8 and 9.

¹ These figures are derived from the normal curve which includes 68.26 per cent between $\pm 1\sigma$, 95.44 per cent between $\pm 2\sigma$, and 99.73 per cent between $\pm 3\sigma$.

ADMINISTRABILITY

Another characteristic that all good measuring instruments have is ease of administrability. Under this category may be included (1) ease of giving and mechanical make-up, and (2) ease of scoring.

EASE OF GIVING

Ease of giving depends upon the adequacy of instructions. Good instructions should be prepared both for the tester and for the subject. Clear-cut directions are necessary for the tester which are beyond those intended for the subject. The tester needs to know the directions for each part of the test, what is the total possible score for each student, and above all, precise time limits. For example, sometimes the tester must read aloud to the subjects while they follow along reading silently. Does the total time allowed include this reading, or does it begin from the time the students actually start their work? The instructions should be so clear on this point that there could be no confusion. Some tests allow only 5, 10, 15, or 20 seconds for a single item. It is very difficult to time these short items correctly unless the tester uses a stop watch. Most recent tests have the instructions which are to be read aloud in heavy print and explanations for the tester in light print. This is a distinct advantage. Furthermore, the general make-up of the test, such as printing and paper, affects the ease of giving. If a word to be defined is supposed to stand out through being printed in bold letters, then *not* to have these bold letters is a distinct disadvantage.

The instructions to subjects should in general be more detailed and explicit the younger they are. However much instruction is needed *to make the problem clear to the subject*, just that much is necessary. Adequate instructions usually include (1) a statement of what is to be done, in clear unmistakable English; (2) one or two illustrations, correctly marked; and (3) an opportunity for the subject to try his hand in doing a simple exercise. Some tests such as the National Intelligence Tests probably went too far in the use of these so-called "fore-exercises," but others have not gone far enough. It is clearly poor procedure for a group of children or students to start out working on problems whose very nature is vague to them. Good testers like to take the time to glance at these fore-exercises to make sure that the subject has them right, before proceeding with the test proper.

In some tests, and perhaps increasingly in the future, detached answer sheets are used. Instructions under these conditions must be carefully and slowly given. With grades below the seventh there is considerable doubt about the efficacy of using detached answer sheets. Most certainly this doubt is increased if the children are not accustomed to being tested, *i.e.*, are not "test-broken."

THE EASE OF SCORING

Anyone can see that under the best conditions a considerable amount of work is going to be required to correct test papers. Especially is this true if there are some items difficult to score because of some slight ambiguity. Objectivity in scoring makes for ease of scoring. If the test is composed of completion items, then the acceptable answers must be clearly listed. A large number of clever devices have been developed which facilitate scoring. Among these, window stencils were among the first to be used. Cutouts on a cardboard permit the scorer to see only the correct items and he has only to count them up. The Clapp-Young self-scoring device uses duplicating paper with each test. Holes are so cut that if the subject gets the item right his cross is registered on the scoring sheet underneath. All that is necessary to score a paper is to count the number of crosses which fall into squares. This procedure makes for both speed and accuracy. Most rapid of all is the new electrically operated scoring machines. About all the corrector has to do after the machine is set up is to copy down the answer. Such a machine scores in a few seconds 150 items with only a small variation in accuracy.

At the present time these machines are very expensive and cannot be owned by the small school. They require also a special type of pencil. Other mechanical arrangements have been tried out, but most are being superseded by this new electrically driven device.

Even the objective short-answer classroom tests may be scored much more easily if the subjects arrange their answers neatly in a vertical column or better still are furnished with answer sheets which are all alike. By placing the correct answers boldly written on strips of cardboard, the scoring is improved in both speed and accuracy. Window stencils may also be easily prepared for this purpose.

INTERPRETATION AND COMPARABILITY

A striking difference exists between a standardized test and one constructed for an ordinary class. This difference consists largely in a difference of opportunity for interpretation. The standardized test would not be so called unless its *norms*, *reliability*, and *validity* had already been determined and published in the manual accompanying the test. Percentile, age, and grade norms are most frequently used. Percentile norms give the scores which marked the percentile points in the group used in the test's standardization. They are easy to interpret and have the advantage of giving reference points at many levels. The standards for the first tests to be constructed consisted only of the medians or 50th percentiles. It is easy to see the importance of having the 5th, 10th, 15th, . . . , 55th, 60th, etc., percentiles as points of

reference. Thus we can say that Sue, who is in college, scores at the 25th percentile among 700 college students.

Some authors have recommended highly that grade and age standards be published for (1) city, (2) town, and (3) rural areas, since these differ among themselves. The rural child is noticeably handicapped on tests dependent upon reading and vocabulary for the scores received. These separate norms are certainly desirable but might impose too hard a task on the test maker. In lieu of this, some facts could be published in the manual indicating the usual differences on this test between rural and city children. There is something to be said in favor of *one norm* from which children deviate because of their unusual environment. *Local norms* established in a single school system are very helpful. Suppose that the fourth, fifth, sixth, and seventh grades of a certain school system scored usually 4.4, 5.3, 6.4, and 7.5 respectively on the Metropolitan Achievement Test at the end of the year and that these scores had become pretty well established. They then could be used as local norms. As a consequence, when the children of a fifth grade under a new teacher scored only 5.2 at the end of the year the administration need not be unduly alarmed. Or again suppose a child transferred from a neighboring state or school scored 4.4 at the beginning of the year. He could be placed immediately among his peers in the fifth grade, while if the administration went by the national norms he would most likely be placed in the fourth grade. *Standard norms and local norms are essential for good interpretation of test scores.*

If the derived scores are used instead of the raw scores which are obtained from the test, a published table should transmute the raw score immediately into a standard score. The most convenient place for this table is at the bottom of the page. In Pintner Intermediate Intelligence Test, Form A, this transmuting table appears at the bottom of the vocabulary test. If a subject scores 13 points the "13" is quickly checked in the table and the standard score, 158, carried forward to the front of the test to be used in deriving M.A. or I.Q.

| Raw score | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|----------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Standard score | 103 | 108 | 113 | 118 | 122 | 126 | 131 | 136 | 140 | 144 | 148 |

| Raw score | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 |
|----------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Standard score | 153 | 158 | 163 | 169 | 176 | 182 | 188 | 196 | 204 | 212 | 219 | 227 |

In high school, norms derived from the number of months a subject has been studied are useful. Thus we may have a three-month norm, a six-month norm, and a nine-month norm. These monthly period

norms could be supplemented by percentile norms at each of the three periods.

Ease of interpretation is also facilitated by having the reliability and validity clearly established and by having really equivalent forms. The manual should state the size of the coefficient of reliability, the number of subjects involved, and the mean and variability of the population used in the standardization of the results. Good manuals also are clear about the validity of the test, both curricular and statistical. In this manner, if a reading test samples closely the reading experiences of children in the fifth grade, let us say, then when children score well on this test we know that they are achieving the objectives which are desired. Equivalent forms aid greatly in interpreting the amount of growth made by children over a designated period of time. They help also when an unsatisfactory test given to an individual child has to be confirmed or denied.

ECONOMY

In most school systems there is great need of economy in administering the testing program. Three types of economy may be mentioned: (1) cost, (2) students' time, and (3) teachers' time. Tests which require as much as 75 cents per pupil are far beyond the funds available for testing in many schools. On the other hand, many of the best group tests may be purchased for 6 to 9 cents apiece. The best tests are not always the most expensive. Some of the more expensive tests are sometimes desirable because their length makes possible the more effective measurement of a complicated objective. Separate answer sheets are also designed both for cheapness and for easy scoring. It is also evident that if too much of a student's time is required for testing too little is left for learning. There is also danger of creating a sullen, negative attitude in students if tests are too long and too involved. In the third place, teachers cannot be expected to stay after school and correct long complex tests. All that was said about the economy of administration applies here. Matters of cost, student time, and teacher time must all be considered in planning for any adequate program of testing.

SUMMARY

Characteristics of a good testing instrument may be divided into five categories: (1) validity, (2) reliability, (3) administrability, (4) interpretability, and (5) economy. Validity is divided into internal or curricular and external. Curricular validity is directly related to the objectives of teaching and attempts to discover types of responses which give expression to the objective, to find a way to quantify them, and to evaluate the responses in terms of the objective. Test con- *

structors in the past have proceeded along practical lines. Items used in educational achievement tests have usually been selected because they are common to several textbooks or courses of study, appear in well-constructed examinations, have proved their social utility, or agree with outcomes of education which a staff of experts has agreed upon. External validity compares the evaluating instrument with other measures of the same objective or outcome. Thus a group intelligence test may be correlated (1) with an individual intelligence test, (2) with a composite of group tests, (3) with teachers' ratings of intelligence, and (4) with school marks. Reliability refers to the accuracy of the instrument, its freedom from chance variation. Four methods of computing it are presented: (1) test-retest, (2) Form A with Form B, (3) the odds-even technique, and (4) the Kuder-Richardson technique. Reliability was shown to depend on the length of the test, the dispersion of the population, and the efficiency of the test's administration. Administrability refers to all those procedures of giving and scoring which affect the efficiency of a test. Instructions for giving and scoring must be clear and unambiguous. Devices for rapid accurate scoring must be furnished. The paper on which the test is printed and the mechanical make-up of the test are also items affecting the administrability of a measuring instrument.

Interpretation depends upon the care with which the norms are established. Age norms, grade norms, and percentile norms are most frequently given. If norms are given in the form of standard scores then transmutation tables should be readily available. Economy of the pupils' time, the teacher's time, and the cost involved are also practical considerations which must be heeded in the selection of any educational measuring instrument.

QUESTIONS AND EXERCISES

1. What is the significance of the question "This test is valid for what?"

2. How have test constructors attempted to secure tests valid in content?

3. Explain and illustrate the relation between (a) social utility and test validity, and (b) psychological and logical analysis and test validity.

4. Illustrate in some detail a test based on psychological analyses. Why is such a procedure difficult to carry out? Is it worth while doing?

5. Explain the function of the criterion in securing test validity. What criteria have been used? Are they satisfactory? Explain.

6. Describe the procedure used by the author in validating group tests of intelligence.

7. How is reliability computed? How can the standard error of a score be looked on as a measure of reliability? Explain. Given a score of 90 with a standard error of 6. Interpret.

8. What factors affect reliability?

9. What is the variability of the population on which the test is standardized of such great importance? What is the best age range to use in standardizing a test? How would the use of this narrow age range affect a test's reliability?

10. How is the coefficient of forecasting efficiency useful in explaining measures of reliability? Illustrate.

11. What functions have fore-exercises in the administration of a test? Explain the need for adequate instructions.

12. How can scoring be made more economical of time?

13. For what purpose are derived scores used?

14. On what factors do interpretation and comparability of a test depend? How can they be made more effective?

BIBLIOGRAPHY

Books

BINGHAM, W. V.: *Aptitudes and Aptitude Testing*, "Selection of Tests," pp. 209-223. New York: Harper & Brothers, 1937.

CRONBACH, LEE J.: *Essentials of Psychological Testing*, pp. 48-83. New York: Harper & Brothers, 1939.

GARRETT, HENRY E.: *Statistics in Psychology and Education*, 3d ed., Chap. XII, pp. 380-403. New York: Longmans, Green & Co., Inc., 1947.

GUILFORD, J. P.: *Psychometric Methods*, pp. 417-418. New York: McGraw-Hill Book Company, Inc., 1936.

HORN, ERNEST, and MAUDE McBROOM: *A Survey of a Course of Study in Reading*, Extension Bulletin No. 93, College of Education Series No. 3, University of Iowa, 1924.

KELLEY, TRUMAN L.: *Statistical Method*. New York: The Macmillan Company, 1923.

REMMERS, H. H., and N. L. GAGE: *Educational Measurement and Evaluation*, Chap. X. New York: Harper & Brothers, 1943.

ROSS, C. C.: *Measurement in Today's Schools*, 2d ed., Chap. III. New York: Prentice-Hall, Inc., 1937.

SMITH, EUGENE R., RALPH W. TYLER, et al.: *Appraising and Recording Student Progress*. New York: Harper & Brothers, 1942.

TERMAN, L. M.: *The Measurement of Intelligence*, p. 55. Boston: Houghton Mifflin Company, 1916.

— and MAUDE A. MERRILL: *Measuring Intelligence*, pp. 9, 12-21. Boston: Houghton Mifflin Company, 1937.

Articles

ALLEN, MILDRED M.: "Relationship between Indices of Intelligence Derived from the Kuhlmann-Anderson Intelligence Tests for Grade I and the Same Test for Grade IV," *Journal of Educational Psychology* (1945) 36:252-256.

BLOOM, BENJAMIN S.: "Test Reliability for What?" *Journal of Educational Psychology* (1942) 33:517-526.

CRONBACH, LEE J.: "Test 'Reliability': Its Meaning and Determination," *Psychometrika* (1947) 12:1-16.

GUILFORD, J. P.: "New Standards for Test Evaluation," *Educational and Psychological Measurement* (1946) 10: 255-282.

GUTTMAN, L.: "A Basis for Analyzing Test-Retest Reliability," *Psychometrika* (1945) 10:255-282.

JORDAN, A. M.: "The Validation of Intelligence Tests," *Journal of Educational Psychology* (1923) 14:348-366, 414-428.

KUDER, G. F., and M. W. RICHARDSON: "The Theory of the Estimation of Test Reliability," *Psychometrika* (1937) 2:151-160.

LANDIS, C., and S. E. KATZ: "Validity of Certain Questions Which Purport to Measure Neurotic Tendencies," *Journal of Applied Psychology* (1934) 18: 343-356.

SCATES, DOUGLAS E.: "Unit Costs in the Administration of a Standardized Test," *Educational Research Bulletin* (1937) 16:38-45.

STARCHE, DANIEL, and E. C. ELLIOTT: "Reliability of Grading High School Work in Mathematics," *School Review* (1913) 21:254-259.

CHAPTER 3

Constructing Achievement Tests

The construction of tests and examinations is important both from the standpoint of understanding the more formal standardized tests and from that of evaluating the results of instruction. The number of informal tests given far exceeds that of the standardized printed variety. One estimate has it that the ordinary teacher gives eight tests of his own to one of the commercial variety. It is consequently of great importance that the classroom teacher know how to check up most efficiently on the educational progress of his pupils.

As was pointed out in Chap. 1, there are at least three aspects of the learning process which throw light on our test construction: (1) the definition of objectives, (2) the provision of the pupils with those experiences whereby the goals or objectives are achieved, and (3) the measurement of the results obtained in order to know to what extent the goals have been reached, the objectives achieved. Each of these procedures modifies the others. If the objectives are reached, then the teacher can be satisfied that his objectives are achievable and that the procedures utilized in collecting and arranging materials by teacher and pupils have been satisfactory. On the other hand, if many of the pupils have not achieved the objectives decided upon, then both procedures and objectives need to be studied and possibly modified. Without this final process of evaluation and measurement, futile objectives and inadequate experiences continue and tend to become hardened into custom.

In this chapter there will be a discussion of the construction of short-answer, easily scorable, objective types of test as well as of the essay type of examination. A complete treatment of these topics with adequate illustrations would require a volume in itself. If the student will master the contents of this chapter and then study the types of test construction used in standardized achievement tests, he will be able to construct satisfactory tests of his own.

CONSTRUCTING CLASSROOM TESTS

The proper construction of classroom tests depends, in the first place, upon a detailed statement of the objectives to be achieved.

The objective agreed upon determines the type of examination to be constructed. In general, objectives include (1) facts, information, and skill; (2) techniques and methods; (3) types of mental processes, such as the capacity to interpret data and to collect and organize it; and (4) certain attitudes, ideals, interests, and values. When these objectives have been carefully defined they must then be analyzed into objectives which can be achieved in a certain length of time. The teacher now has to decide which type of test most nearly indicates the achievement of that objective. If a number of good test items could be formulated as the learning takes place, much strain and effort would be saved near the end of the course and more effective evaluating instruments would be constructed.

ESSAY-TYPE QUESTIONS

Theoretically, for a student to gather his thoughts from a well-stocked memory, sift them out, and apply them intelligently to the topic at hand is to display most effectively his educational attainments. Such an answer would be in response usually to an essay-type question introduced by "discuss," "describe," "explain," "compare," or "indicate." Had there been substantial agreements among those who attempted to score such attempts on the part of the student, there probably would not have arisen the movement for short-answer questions.

One of the clearest cases of the weakness of the essay type of examination occurred in a study in England.¹ This case is especially noteworthy because those who graded the examinations were expert graders whose main business in life was allotting marks to papers sent in to a central office. The same 48 English papers were graded independently by seven of these graders, with the following results:

| Examiner | Fail | Pass | Credit | Special credit |
|----------|------|------|--------|----------------|
| A | 1 | 16 | 27 | 4 |
| B | 0 | 2 | 34 | 12 |
| C | 7 | 30 | 11 | 0 |
| D | 0 | 9 | 36 | 3 |
| E | 5 | 16 | 27 | 0 |
| F | 2 | 7 | 37 | 2 |
| G | 19 | 12 | 17 | 0 |

Note, if you will, the difference in the number of failures allotted by these experts. While G fails 19, B and D fail not a single paper. At the

¹ Hartog, Sir Philip, and E. C. Rhodes, *An Examination of Examinations*, p. 20. New York: The Macmillan Company, 1935.

other end of the scale B gives 12 special credits out of the 48 papers, while equally competent C, E, and G give none at all. Look again at B and C. B's marks lean heavily toward the higher end; C's toward the lower end. It is thus clearly seen that the mark a paper receives depends significantly upon the grader into whose hands it falls.

There are certain surmountable difficulties in the essay-type question which must be met if the agreement of the graders is to be increased. Among these are the different values placed upon certain aspects of topics, disagreements about counting off for misspelled words, and oddities of grammar which can be provided for by consultation among the graders. Certain other difficulties are more difficult to overcome. Four of these more serious ones are the following:

1. That type of answer to essay questions known as *padding* is usually made up of gleanings from general reading and conversation. These rather glittering generalities may be woven together into a fabric composed of truth, half-truth, and downright error. How should such a discussion be judged—fail, poor, fair, or average? There is no doubt that considerable credit is sometimes given for just such an answer.

2. The discussion takes a direction not contemplated by the constructor of the test. A student may honestly interpret a question to be answered in one way while its writer intended it to be answered in another. This may be due to the lack of precision exercised in the item's construction. Suppose the answer is undeniably good, though not in the direction intended—how should it be graded? Further discussion concerning the manner in which the essay question itself may be improved appears on pages 59 to 63.

3. The grammar is satisfactory but there is a lack of logic in its presentation. There are students whose ingenuity in tangling up the logical arrangement of a test is truly a masterpiece. By the side of one question with clear cut sequence and excellent integration, will appear an answer with almost no sequential arrangement, no discoverable logic, and yet much of the material presented is factually correct.

4. In the essay type of examination or test, the sample is compelled to be rather narrow. Four or five topics out of the 15 or 20 are about all that can be well discussed in a test of 2 hours. The student may be better prepared on the topic not discussed than on the one included in the test. For these reasons the conscientious teacher who truly desires that marks and grades be genuine indicators of educational achievement finds himself frustrated.

Because of the reasons just described, essay-type questions and examinations have generally shown both low reliability and low validity. Teachers of the same subject were unable to agree on a mark for a

single paper photographed and sent to them. For example, in one of the earliest studies a photostatic copy of a geometry paper was sent to 116 high school mathematic's teachers to be graded.¹ The scores ranged from 28 to 92. This case is doubly interesting because there was no padding, no misunderstanding of the question, and no particular problem of counting off for poor spelling or bad grammar. These procedures were repeated in English, social science, and other subjects. In the second place, wide variations in the percentages allocated to each school mark frequently occurred even in the same school, so that the number of failing marks varied from 0 to 15 or 20 per cent while there was a corresponding variation in the percentages allocated to other marks. It seemed that the mark a student received depended almost as much upon the instructor he had fallen heir to as upon the progress toward the defined objective. As a result of many such studies of unfortunate experiences with ordinary school tests and examinations, there was a rather rapid development of short-answer questions.²

SHORT-ANSWER QUESTIONS

Short-answer questions are intended to be framed in such a way that the crux of the matter, the base on which the whole answer rests, is the answer to the item. In translating a sentence from a foreign language into English the exact translation sometimes turns on the meaning of one word. Could the meaning of this word be discovered, the jam would be broken and the thought flow on without interruption. If this word is not known the translation is limping and ineffectual. The builder of short-answer tests welcomes such a word. He embodies it in a multiple-choice test or in a completion test. Now he no longer has to try to disentangle the translation of the whole paragraph. In this case the correct answer may be achieved without padding, without new direction to the discussion, without logical difficulties and without inadequate sampling. Sampling can be more satisfactory because many more items can be included than would be possible in the essay examination. There are two general types of short-answer testing: (1) those based on recall, and (2) those based on recognition.

SHORT-ANSWER TESTS BASED ON RECALL

Two types of tests based on recall are (1) simple recall, and (2) completion.

¹ Starch, Daniel, and Edward C. Elliott, "Reliability of Grading Work in Mathematics," *School Review* (1913) 21:254-259.

² See the rather complete discussion in Ross, C. C., *Measurement in Today's Schools*, 2d ed., pp. 44-49. New York: Prentice-Hall, Inc., 1947.

Simple Recall

One of the oldest methods of attempting to objectify the responses to tests and examination is that of simple recall. This procedure differs from the usual essay type of question by limiting the answer to one word or one phrase. Indeed, most complicated questions involving explanation or discussion may be broken down into several questions with short answers. Care must be taken to phrase the questions in such a way that the answer is definite and short—a single word if possible. Of course, such brevity and precision require that the subject matter be of a definite nature also.

The blanks for the answers, long enough for legible writing in all cases, should be placed in a vertical column to the right of the question. It is most important that all the acceptable answers should be listed on the scoring sheet. For testing understanding rather than rote memory, questions and statements should be expressed in language different from that used in the textbook. In general, usage dictates that one point be given for each item correct.

Illustrations

The items of the test may be expressed (1) in the form of a question, (2) in the form of a statement, or (3) in the form of a stimulus word.

1. Questions

- | | |
|--|----------|
| a. In the expression "Dear Sir" at the beginning of a letter what punctuation follows "Sir"? | a. _____ |
| b. What is the scientific name for the splitting which occurs in the uranium atom in the formation of atomic energy? | b. _____ |
| c. What is the name of the port on the Adriatic that is claimed by both Italy and Yugoslavia? | c. _____ |
| d. If M.A. is divided by C.A. what is the quotient usually called in psychology? | d. _____ |
| e. What is the logarithm of 100 to the base 10? | e. _____ |
| f. What is the grammatical name of a verb when used as a noun? | f. _____ |

2. Statements

- | | |
|--|----------|
| a. Name the outstanding characteristic of the paintings of George Inness. | a. _____ |
| b. Name the president under whose direction the Louisiana Purchase was made. | b. _____ |
| c. Write the future tense, first person singular of <i>aller</i> . | c. _____ |
| d & e. State the names of two men responsible for the theorem on which Thorndike constructed his first scientific educational scale. | d. _____ |
| | e. _____ |

- f. The sum of two numbers is 14. The difference between the squares of the two numbers is 28. Find the larger number. f. _____
- g. Give the number of the amendment to the constitution of the U.S. which was responsible for national prohibition. g. _____

3. Word-or-phrase Form

a. Scientist's

| Name | Most Famous Contribution |
|-----------------------|--------------------------|
| 1. Urey | 1. _____ |
| 2. Arthur Compton | 2. _____ |
| 3. Einstein | 3. _____ |
| 4. Priestly | 4. _____ |
| 5. Thomas Hunt Morgan | 5. _____ |

b. English to French

| | |
|--|----------|
| 1. Chair | 1. _____ |
| 2. Glass | 2. _____ |
| 3. Go (present tense, third person singular) | 3. _____ |
| 4. Hat | 4. _____ |
| 5. Desk | 5. _____ |

The simple-recall type has both advantages and disadvantages as compared with other short-answer techniques.

Advantages

In the true-false, multiple-choice, etc., types the correct answer demands only the recognition of the right answer. This factor introduces the unreliability involved in guessing, as in the multiple-choice technique wherein the process of eliminating the answers that are certainly wrong leaves the choice to be made from two items rather than from five as was intended. Now simple recall, since there is no recognition to be made but only recall, reduces the process of guessing to a minimum. One can be assured that the path followed in the solution of the problem is pretty largely controlled. As compared with the usual essay type it directs the thought process toward a definite goal and prevents padding and bluffing. The form is one frequently used and hence is familiar to the subject. Finally, it is fairly economical of space, is easy to construct, and allows a wide sampling of subject matter in a comparatively short time.

Disadvantages

As compared with the recognition type of short-answer test this one is harder to score because it is impossible to predict beforehand exactly the answer which the subject will give. Some of these deviate slightly from the lists of acceptable answers furnished by the key and

hence are difficult to score. The more precise the form of the question, the less does this difficulty appear. In these days of scoring done by automatic machines this interpretative aspect of the answer is a distinct drawback. It is not, however, a drawback to the ordinary teacher trying honestly to evaluate the progress his students are making in the area of instruction. Probably the greatest disadvantages of this type arises in the difficulty of making up items which call for the higher types of mental processes. Naming, citing, giving the author, or his works, are closely related to rote memory. In mathematics and science, it is easy to overcome this difficulty, a fact easily demonstrated by noting that problems in arithmetic and algebra fall naturally into this form.

Sentence Completion

Most of the characteristics listed under "simple recall" also belong under "sentence completion." Sentences, from which certain words are deleted, should be definite and clear and should be composed anew and not simply copied from the book. Then, too, all possible answers for each blank should be carefully listed so that the scoring can be objective. Most difficult of all in constructing the sentence-completion test is to achieve that nicety of balance between supplying just enough material to make the solution possible and giving so much information that the answer can be guessed at. Blanks may occur at any place in the sentence and should possess three characteristics: (1) they all should be of the same length, (2) all should be numbered, and (3) the correspondingly numbered blanks should be placed in a vertical column to the right or left of the sentences. Blanks should almost always call for but one word. In general the larger the number of blanks the more difficult is the item. If the blank is placed at the end of the sentence the completion sentence becomes a simple recall. Here are a few examples:

- | | |
|---|----------|
| Heredity is the (1) relation between (2) generations. | 1. _____ |
| The colored part of the eye, called the iris, (3) or (4) as the | 2. _____ |
| amount of light increases or (5) . | 3. _____ |
| | 4. _____ |
| | 5. _____ |
| The chromosome is an (1) of (2) threadlike (3) . | 1. _____ |
| | 2. _____ |
| The coefficient of correlation represents the (1) degree of (2) ex- | 1. _____ |
| isting between (3) traits in the same (4) of individuals, each | 2. _____ |
| individual being measured (5) . | 3. _____ |
| | 4. _____ |
| | 5. _____ |

The completion sentence probably finds its greatest usefulness in testing the development of a rather complex idea in a whole paragraph. Used in this manner it approaches very closely an instrument for validating the higher thought processes.

Its advantages and disadvantages are nearly the same as those of the simple-recall type.

SHORT-ANSWER TESTS BASED ON RECOGNITION

Four types of short-answer tests based on the capacity of the individual to recognize the correct answer among several presented will be described. They are (1) multiple-choice tests, (2) true-false tests, (3) matching tests, and (4) tests of the higher mental processes.

Multiple Choice

In the type of short-answer test known as *multiple choice* the right answer to a question appears among a number (usually two to four) of wrong ones. Unless these wrong ones are as plausible as the correct one, the purpose of the test is defeated. The answers that are not plausible are immediately eliminated from the test, and the subject can then make his choice between those left. Plausible wrong answers, called *distractors*, are sometimes secured by giving the item as a short-answer test in a preliminary tryout. Pupils themselves will write the wrong answers which may then be used as the wrong alternatives. Illustrations of this occurred in preparing tests for the selection of good navigators in the Army Air Force. The item was first given in the incomplete form. The errors were then compiled and the four most frequent errors were used as alternates in the multiple-choice test. Sometimes distractors are used which lead the ignorant to the wrong answer. If these distractors have a logical or bookish connotation they work better. Two illustrations from Inglis Tests of English Vocabulary¹ illustrate the use of distractors:

A regular *hexagon*. (1) six-sided figure. (2) old witch. (3) model. (4) nuisance. (5) assembly.

Answer No. 2, "old witch," is associated with the word "hex," to bewitch. A second illustration:

It is a result of *collusion*. (1) bumping. (2) conflict. (3) kindness. (4) fraud. (5) lawlessness.

"Collusion" might be mistaken for "collision," and hence "bumping" would be a distractor.

¹ Boston: Ginn & Company. By permission.

Another factor that adds to the plausibility of the alternates in the item is their homogeneity. The more like each other the alternates are, the harder they are to distinguish and the finer is the discrimination required. An illustration from the Columbia Research Bureau American History Test¹ shows exactly what is meant:

Americanization is the process of—

(1) Keeping foreigners out of America (2) extending American trade by means of subsidies (3) teaching American ideals to foreigners (4) becoming naturalized (5) protecting American industries. (—)

All answers are plausible and hard to distinguish unless one knows the answer.

This illustration about "Americanization" also shows (1) that there is no punctuation except the period at the end of the statement, (2) that parallel construction is maintained in all the items, and (3) that the multiple-choice technique is better than that of simple recall when the answer to a question might be long and complex, or when the answer might be given in one or two different ways. Moreover the three illustrations use the statement form rather than the question form. Had this latter form been used the item would have read "What is the process of Americanization?" and the alternates might then have been introduced by "It consists of ." Some test makers prefer this direct question form because, they say, "it is easier to construct; it is in a form with which the pupil has had experience and is less likely to contain cues to the correct answer." These preferences seem to be largely matters of opinion except in the naturalness of the question form to children in school.

The conditions under which the multiple-choice form is to be preferred to that of simple recall were indicated. Under certain conditions, on the other hand, the simple-recall form is to be preferred to that of multiple choice. When the answer is a number or symbol the simple-recall form is best. Sometimes, too, try as one may he cannot find more than two plausible choices for a certain item. Under these conditions, also, the simple recall is better. Arrangement and punctuation of items need some consideration. In arranging items of the multiple-choice form, care must be exercised not to use a regular cycle of answers. Each of the positions (1,2,3,4,5) should be used about the same number of times, but there should be no logical arrangement of items. Use no punctuation between choices but simply skip three spaces. Place the proper punctuation at the end. *Place parentheses around the numbers which are just in front of the possible answers if the answers are numbers, otherwise not. Here is an example:*

¹ Yonkers, N.Y.: World Book Company. By permission.

One should send a check by:

- 1 first class mail 2 express 3 parcel post.

But note parentheses in the following example:¹

Banks usually pay from (1) 1% to 2% (2) 2% to 5%.

The reader will notice that this principle has been violated in some of the previously quoted tests.

Advantages

The multiple-choice form is the most flexible of all the forms of short-answer tests. Its alternates may be so near together in meaning that it takes much keenness of discrimination to distinguish between them, or again they may test simply information acquired by rote. They need not be corrected for chance if there are more than two alternates. They are to be preferred to simple recall in complicated ambiguous problems of some length. The reliability of this form is high.

Criticisms

Multiple-choice items are difficult to construct. It takes as much time to construct one good multiple-choice item as to construct three to four simple-recall or true-false items, and they occupy as much space on the page. Plausible alternatives are hard to find. It also takes more of the pupils' time to answer multiple-choice items than to answer true-false items. A great impetus has been given the use of the multiple-choice form since the advent of the IBM scoring machine. This machine scores a whole test accurately provided the answers are placed in certain defined positions. The multiple-choice form with its five positions lends itself admirably to machine scoring.

True or False

In the constant-alternatives form of the short-answer questions the pupil is asked to render a judgment about the statement as a whole. In the vast majority of cases the judgment rendered is whether the statement is true or false, and hence this form is most commonly known as the *true-false form*. It can be used in almost any field of learning and to evaluate the materials as well as the mental processes involved. A few samples follow:

- T or F 1. The constitution of the U.S. safeguards continuity of the Senate by specifying that only one-third of the Senators shall come up for election during one year.
- T or F 2. The Mississippi River is usually regarded as the Great Divide.

¹ Rinsland, Henry D., *Constructing Tests and Grading*, p. 47. New York: Prentice-Hall, Inc., 1937.

- T or F 3. Two triangles are equal if a side and an angle of one equal the corresponding side and angle in another.
- T or F 4. Two root words entering into the conjugating of the French verb *aller* are *vado* and *eo*.

When properly constructed, this short-answer form can be made to sample a very large number of items in a short time. It is comparatively easy to construct and to score, although its scoring is subject to a few more errors than is the case with other forms.

Suggestions for Improving the Construction of True or False Items

First of all, the statements should not be lifted bodily from a textbook with perhaps a slight change in the wording to make some of them false. It is much better to have the idea embedded in a fresh array of words. The language of the items should be within the comprehension of those taking the test. Moreover, the statement should be clear and unambiguous, not clouded in meaning by too many qualifying clauses or double negatives. Whenever possible, quantitative statements are better than "more" or "less" or other indicators of comparisons. For example, a statement such as

- T or F Foster children adopted into good homes increase their scores on an intelligence test.

may be improved by changing it to

- T or F Foster children adopted into good homes increase their I.Q. scores 6-7 points on the average.

Certain determiners of a statement's truth or falsity are to be avoided. Sentences using such determiners as "totally," "entirely," "completely," "solely," "absolutely," "always," "never," "only," "alone," and such other words which imply universals are usually false. On the contrary, sentences using "should," "may," "most," "some," "often," are more than apt to be true ones. By actual count of words it has been shown that long sentences (with more than 20 words) are likely to be true. These are the principal suggestions, although there are some lesser ones.

The constant alternatives may appear as T or F, yes or no, T or 0, or by a slight addition, T or F or ?. This last one, "true, false, or question" allows some leeway in testing those principles which we sometimes answer by saying "That depends." Still another variant permits further shades of belief:

- T, f, Ut, Uf The mean derived from grouped data is dependent for its accuracy upon the distribution of the items within the interval.

To such an item an individual may respond T = true, f = false, U_t = usually true, or U_f = usually false. In this form it approaches that of multiple choice. Still another modification is introduced by instructing the pupil to correct with one word each false statement but to do nothing further to the true statements.

| | | |
|----------|--|-------------------------|
| T or (f) | Montgomery was commander-in-chief of the allied armies. | 1 (<i>Eisenhower</i>) |
| (T) or f | One of the greatest landings of men in world history took place on the Normandy beaches. | 2 (———) |

This procedure of correcting the false statements is better liked by pupils and students.

Finally, one other suggestion concerning the arrangement of true or false sentences is in order: the sequence must not be a logical one. Such a sequence, for example, as T, f, f, T followed by T, f, f, T would soon be detected and the sentences thereafter marked correctly by reason of the student's ingenuity in discovering the logic of their arrangement and not because of his understanding or information. To avoid any semblance of logical arrangement one may be governed by the toss of a coin. For example, toss a coin and keep the record of heads and tails. Whenever a head falls, make the statement true; whenever a tail falls, make it false.

Correcting True or False Statements

It is easy to see that when there are only two alternatives a pupil has a 50-50 chance of getting an item correct. To avoid the influence of chance in the score, many students have recommended that the score be obtained from either of the following formulas:

1. $S = R - W$. Score equals the number right minus the number wrong. Thus, if there were 100 items, of which 50 were correct and 50 wrong, the score would be exactly 0. The omitted items do not count in either direction. In scoring a true-false test one may make a cross for the wrong item and a short line for those omitted. If we draw a heavy line under the last item attempted by the subject and then secure the total attempted by glancing at the number of the last item attempted, the following formula gives the same score as the right minus wrong one.

2. $S = T - O - 2W$. Score equals "total" (as indicated by the number of the last item attempted) minus the number omitted, minus twice the number wrong. The number omitted would then mean the number omitted up to the last item attempted (not those the subject had not tried at all). The symbol W would mean those wrong, as before. Suppose there were a test of 40 items. The 35th had been the last one

tried but the 25th, 28th, and 34th had been omitted. Four items were wrong. Under these conditions, using Formula 1, $S = R - W$, the score is $28 - 4$, or 24. If we use Formula 2, $S = T - O - 2N$, we get $S = 35 - 3 - 2(4)$, or 24. Formula 2 is the more practical formula, since for the most part a score can be secured by multiplying the number of wrongs by two and subtracting this number from the total attempted.

This whole matter of correcting for chance has come under minute scrutiny. Such formulas are based on a very large number of draws of samples. If there were 2,000 items these formulas would be quite satisfactory, but with 100 items chance sometimes acts very queerly. Who has not drawn good hand after good hand in an evening of bridge while at other times the opposite is true? Affecting directly the scoring are the directions given. Shall we say to the subjects, "You have plenty of time and I want you to answer all items. If you aren't sure, guess," or shall we say to them, "Mark only those items you are certain of. Do not mark those of which you are not certain"? The second set of instructions on its face seems the most sensible. But individuals differ so greatly in their carrying out of these instructions. The quiet precise individual may not try more than 25 out of 40, all of which will be right. Another more venturesome lad will try 35 and make three mistakes. The former student receives a score of 25; the second, 29. Under these conditions the formula correcting for chance would necessarily be applied although the awareness of its limitations in correcting chance in a small number of items would be apparent to all. The instructions to the subject to answer all the items seems about as fair as any other. If the subject disliked guessing very much he could disregard the instructions. If students try all the items when they are instructed to do so the correction for chance errors may be omitted since the uncorrected score correlates perfectly with the corrected one.

To score many columns of T's or F's is very fatiguing to the eye. A key made of stiff cardboard with perforated holes through which all the true scores may be seen at a glance is a very helpful device.

Matching

In constructing a matching exercise two procedures may be followed. In the first one, called *completion matching*, an essential word or phrase is omitted within each sentence of a list of sentences. At the end of these sentences is a list of words or phrases which contains the best answer for each of the omitted words. This form differs from the sentence-completion form in that in the completion-matching form there may be 10 of 12 answers in a column from which the correct completion to, say, 8 or 10 word omissions may be made. If the sentence-

completion form had been used there would have been four or five possible answers for each sentence, or 40 to 50 altogether. From Rinsland¹ appears this sample:

| <i>Part I</i> | <i>Part II</i> |
|--|-----------------|
| 1. () The number to be multiplied is the (). | 1. difference |
| 2. () The result of addition is called the (). | 2. dividend |
| 3. () The number to be divided is the (). | 3. divisor |
| 4. () The result of subtraction is called (). | 4. minuend |
| 5. () The result of multiplication is called (). | 5. multiplicand |
| | 6. multiplier |
| | 7. product |
| | 8. sum |

In the second type, called *column matching*, two columns of statements are placed side by side and then the numbers of one column are matched with the numbers or letters of the other. An example from genetics follows:

One of the statements in Part I defines, illustrates or in some other way belongs with the items in Part II. Place the correct number from Part II in front of the appropriate letter in Part I.

| <i>Part I</i> | <i>Part II</i> |
|---|-----------------------|
| () a. Occurs in all individuals during the first generation | 1. acquired character |
| () b. Only one of two alternative characters resides in the germ cell. | 2. congenital |
| () c. Red green colorblindness follows the defective X-chromosome. | 3. dominant |
| () d. Transmitted through the germ cells. | 4. inherited |
| () e. Appears in the ratio of 1 to 3 in the second generation. | 5. instinct |
| () f. The acquisition of a disease from the mother during the embryonic state. | 6. purity of gametes |
| | 7. recessive |
| | 8. sex-linked |

Some of the characteristics of matching may be observed in these two illustrations. The more obvious ones have to do with form. There should be more answers in Part II than are needed in Part I. This reduces the matter of chance to a minimum. Only one answer must be correct. It helps the subject if the items in Part II are arranged alphabetically or logically. Great care should be taken to avoid having any clues in Part I or Part II which would suggest the answer, such as both

¹ Rinsland, Henry D., *Constructing Tests and Grading*, p. 104. New York: Prentice Hall, Inc., 1937. By permission.

singular and plural forms of words. Sometimes the connection is suggested through identity of singular subject and singular verb or vice versa. Generally speaking, 7 to 10 items in Part I and 10 to 12 in Part II would be about as many as would be practicable. It is clear also that all the items of Part I and of Part II should be on the same page.

Less obvious than the just-mentioned characteristics is that of homogeneity. All the items in Part I should be like each other, *i.e.*, homogeneous. The elimination of guessing may be greatly facilitated by the homogeneity of the items. All items of Part I of the first illustration could be subsumed under the four fundamental arithmetic operations; while the items of Part I in the second illustration can be placed under inheritance. The more homogeneous the items the more difficult to guess the answer correctly. Hence the dictum: if you wish to make the items more difficult make them more like each other.

There are a large variety of relations to which matching is applicable: cause and effect, dates and events, authors and their writings, diagrams and charts, principles and their illustrations, inventions and inventors, angles and their names, tools and their uses, names of compounds and their chemical formulas, and many others.

Advantages

Many questions can be answered in a short space because the same set of answers can be used for a large number of items. Guessing is reduced under the usual method of construction but may be reduced to a minimum by having several items use the same answers. Its greatest usefulness comes in answering questions *who*, *when*, *what*, and *where*. Whether or not it tests the more complicated mental processes depends upon its construction. By matching principles and their illustrations the subject is called upon to discriminate, compare, and conclude. Such a procedure calls for the same sort of mental processes which are demanded when an individual is asked to give an original illustration of a principle he has learned. This type of short-answer test is capable of making a rapid survey of a particular phase of a subject-matter area.

Disadvantages

Matching tests are difficult to construct. It is so easy to leave undone the large variety of specifics which need to be heeded in constructing them. Clues that one had never suspected and more than one correct answer are apt to appear most unexpectedly. Furthermore, it fits so well simpler items such as events and their dates that more complicated associations are apt to be neglected. Small units of subject matter rarely furnish that homogeneity demanded of a good matching test

and hence a small unit of instruction is difficult to test adequately by using this form.

SHORT ANSWER TESTS: HIGHER MENTAL PROCESSES

Thus far in our discussion of the construction of short-answer tests no special emphasis has been placed on testing the higher mental processes. It is believed, however, that such processes may be brought into play in answering true-false, completion, simple-recall, multiple-choice, or matching questions. It is the purpose of this section to call attention to the possibilities of evaluating the capacities of individuals (1) to interpret new data which are presented, and (2) to apply principles learned to new situations. One might even like to measure the understanding of the nature of proof itself, but thus far such small progress has been made in perfecting instruments for that undertaking that this topic is omitted in the present discussion.

Ideally, it would be best to check the whole process of observing, guessing, formulating hypotheses, gathering data, and finally making inferences and other interpretations from the data gathered. So long is this process and so few are they who are called upon to carry it through that no objective criteria have been formulated which can be applied at all stages of the total process. It, however, has been found practicable to set up procedures by which the capacity of an individual to interpret data already collected by others can be evaluated both as to the type of conclusion reached and as to the manner in which the judgment was achieved. For a discussion and illustration of an attempt to analyze and measure clear thinking, see the discussion on pages 18 to 21.

In one volume¹ attempts were also made to develop tests for principles of logical reasoning and for the nature of proof. The reliabilities of all these instruments were in the neighborhood of .90 as calculated by the Kuder-Richardson formula. In general, this formula gives a slightly lower coefficient than other procedures for computing reliability. As a whole these procedures for testing specifically the higher mental processes are still in the experimental stage. The importance of clear thinking makes experimentation in this area extremely worth while.

There are several other methods of constructing short-answer items such as analogies, classification, rearrangement, and cause and effect. Many of these may be observed in the chapters on intelligence tests and in our treatment of personality inventories. Most of them are variants from the types introduced in this chapter.

¹ Smith, Eugene R., Ralph W. Tyler, *et al.*, *Appraising and Recording Student Progress*, pp. 111-124. New York: Harper & Brothers, 1942.

ORGANIZATION AND ARRANGEMENT OF TESTS

Let us now assume that the objectives have been defined, and items for the construction of the test have been accumulated. Let us assume further that the items have been carefully edited and cast into the most desirable test form. There still remains the organization of the items and their arrangement.

Assemble the items under test forms. Suppose, for example, that the course had been a survey of American history covering the period from 1865 to 1900 and that some of the items were true-false, some of them simple-recall, and some others matching. In the arrangement of three or four forms on one topic there would necessarily be a small number of true-false items, a small number of matching items, and a small number of simple-recall items. By assembling all true-false items, simple-recall items, etc., into one division of the test the same set prevails for a much longer time and the confusion of shifting mental sets is avoided. For this reason, it is better to place *all the true false items in one section, all the matchings in another, and all the simple-recall items in still another.*

Arrange the items from easy to hard. In general it is better to arrange items roughly from the easy to the more difficult. An exact grading of items according to difficulty is manifestly impossible until they have been tried out with a number of subjects and the percentage passing each item calculated. The teacher from his acquaintance with the class and the difficulty of the items can arrange them into four or five groups of increasing difficulty. If the difficult items are placed first the subjects may spend so much time on the first items that there is no time left for the easy ones which come later or else may get so discouraged because they seemingly cannot answer enough items to pass the test that they give up completely. The easy-to-hard arrangement gives the subject confidence, and if he takes too much time on the more difficult items he has at least finished the major part of the test. The items should range in difficulty from those almost all the class get right to items difficult enough so that very few get them right. It is best if the average of the class lies somewhat near half the number of items. This idea should be kept in mind by the test constructor, but a mean lying between 35 and 65 in 100 items would not greatly disturb the efficiency of the test.

Arrange items so that their answers cannot be guessed or worked out logically. Suggestions have already been made how this is done in true-false, matching, and multiple-choice tests. Some sort of chance arrangement is best. In the multiple-choice form see that each of the positions (1,2,3,4,5) has the correct answer about the same number of times.

The tester must be provided with enough extra pencils so that there will be no delay in the progress of the examination. He must read over the instructions aloud with the children, answer their legitimate questions, and make sure that they understand exactly what they are to do before they begin. Children should be practiced in a preliminary way before they take these short-answer types of test. The tester must see that the children are not disturbed and that the stop signals are properly given.

Arrange numbered vertical columns on either the right or left so that the responses can be easily scored. These empty dotted lines must be long enough in completion and simple recall to write the answers. Little children especially have a tendency to write larger than do adults. Most authors recommend that these columns be placed to the left of the numbered item in true-false and matching tests and to the right in multiple-choice, completion, and short-answer tests.

Score tests by using a prepared key. If the instructions concerning the correct placement of the answers are carried out, one may then make a good scoring sheet by writing in the correct answers with a red pencil. Place these filled-in sheets right by the answers of the subjects and checking may go on at a rapid pace. If there is a large class it may pay the grader to paste the answers on a cardboard strip which holds its position without bending too much. True-false corrections are apt to cause some trouble. If "T or f" is placed just to the left of each item, then the correct items may be punched out of a cardboard with a circular punch. If this is then placed over the score column the correct items may be seen through the holes.

Give children detailed instructions. Generally speaking, blanks will be left on the outside of the paper for the date, the name, and the grade and for both part scores and the total score. Explain to the subjects exactly what is to be done in each case. Explicit instructions must also be given to the children concerning the following of directions, whether they may ask any questions, and whether they may use any leftover time to work on tests already finished. Especially is it important to explain to children the effect of guessing in the true-false test. If they are to be penalized for the wrong answers, then they should be told about it. Many investigators prefer that the children be asked to go all the way through the true-false items and mark those they know, then go through the items a second time and guess at the rest of the items. In this case the items need not be corrected for guessing.

IMPROVING THE ESSAY TYPE OF EXAMINATION

In evaluating any type of examination the first consideration is its effectiveness in the measurement of the objectives decided upon at the beginning of the course. Unless the objectives aimed at in the course are

tested by the type of examination used, the examination is necessarily useless. To be more specific, essay questions frequently ask the student to "compare," "contrast," or "discuss." The adequate answering of such questions depends on the manner in which the course has been taught. All along throughout the course the student must have practice in comparing, contrasting, and discussing. He must know beyond a doubt that "compare" means to set down facts or lines of evidence side by side and from their contemplation come to a reasoned conclusion. And thus it is with "contrast" and "discuss." It is impossible for students to make such contrasts, comparisons, etc., unless they have been trained in the forms and procedures used to arrive at reasoned conclusions. When such conditions have been met the essay type of examination gains in altitude because it requires the students to perform complex mental processes involving comparison and inference.

It is very probable that the higher mental processes involved in comparing, contrasting, and discussing can be appraised by the essay question just as accurately and with greater economy than by the objective types of testing described on page 20. Whereas two or three pages with a variety of possible inferences are needed to construct an objective test only a short sentence asking for the precise comparisons and inferences may be all that is needed for the essay type of examination.

For these reasons it is of the first importance for the teacher to know how (1) to construct effective essay-type questions, and (2) to score them more precisely so that the reliability of the examination based on them will be adequate.

VALUE OF THE ESSAY-TYPE QUESTION

The limitations and undoubted weaknesses of the ordinary essay-type questions and of the examinations composed of them have already been described on pages 41 to 43. In spite of these criticisms there were those who felt sure that this type of question was valuable because it assayed many of the higher mental processes involved in the organization and evaluation of experience. It was and has always been the only medium used in writing compositions and preparing articles in journalism courses. It has, on the other hand, been woefully misused when it inquired for details of information which could have been secured much more effectively by the short-answer tests such as the true-false, short-answer, multiple-choice, etc.

A historian, A. C. Krey, who is greatly interested in the teaching as well as the testing of the outcomes of social science, writes as follows:¹

¹ Kelley, Truman L., and A. C. Krey, *Tests and Measurements in the Social Sciences*, p. 480. New York: Charles Scribner's Sons, 1934. By permission.

Furthermore, such minute sampling of social science knowledge [by means of short-answer tests] clearly did not constitute a test of the student's comprehensive knowledge, or of his ability to develop sustained exposition of large ideas and to include the conditional elements which qualify any but the most simple of social situations. In other words, the extremely short answer form of the test seemed an artificial limitation which must confine such tests to the measurement of only the fragmentary beginnings of social science knowledge.

It is possible through essay questions "to develop sustained exposition of large ideas and to include the conditional elements which qualify any but the most simple of social situations." When items selected from a large number are to be brought to bear in a central topic, when they are to be compared and evaluated, and from this procedure an inference is to be drawn the essay question is more effective than the short-answer type. For these reasons, the essay question has weathered the storm of criticism.

It is the purpose of this discussion to show how (1) the questions can be so improved as to register more precisely the desired processes, and (2) the accuracy of scoring can be greatly increased by deciding on the items to be counted before the tests are scored and by instructing the students in the essentials of good answers before they begin the test.

IMPROVING QUESTIONS OF THE ESSAY TYPE

Substantial progress in describing and illustrating a rich variety of types of essay questions, 20 in all, was achieved by the publication of Monroe and Carter¹ in 1923. Ten years later, Weidemann's² 11 different types of usable questions were made available. From these two lists the author has selected and illustrated 10 different types of questions. It is important to understand that these are simply illustrations, which need to be adapted to the framework of the course which is being conducted.

1. Interpretation

- a. Cite and interpret the following lines of evidence bearing on the problem of maturation in young children: (1) neurological, (2) co-twin control, (3) parallel groups.

¹ Monroe, Walter S., and Ralph E. Carter, *The Use of Different Types of Thought Questions in Secondary Schools and Their Relative Difficulty for Students*, Bureau of Educational Research Bulletin No. 14, University of Illinois, 1923.

² Weidemann, C. C., "Written Examination Procedures," *Phi Delta Kappan* (1933) 16:78-83.

- b. How do you interpret such evidence as "Not a cough in a carload," "Doctors say there is no throat irritation from smoking brand x," etc., when used in radio advertising?
2. Criticism and evaluation
 - a. Criticize and evaluate the effect of the Yalta Agreement.
 - b. Criticize the notion of "independent unit" in heredity.
3. Statement of purpose
 - a. What was Shakespeare's purpose in introducing the witches into *Macbeth*?
 - b. What is the purpose of local government?
4. "How" questions
 - a. How would you set up an experiment to demonstrate the influence of air pressure on the lifting power of a pump?
 - b. How is it possible for an airplane to rise and remain in the air for certain periods of time?
5. Cause and effect
 - a. What was the effect of the removal of price controls on the cost of ordinary commodities?
 - b. What is the effect of high mountains near the coast and prevailing winds on the amount of rainfall in the interior?
6. Statement of relationship
 - a. In what ways is the reliability of a test related to its validity?
 - b. What is the relation between rainfall and crop yield?
7. Comparison and contrast
 - a. Compare the actions of Lady Macbeth with those of Macbeth when they were contemplating the death of Duncan.
 - b. Point out the leading differences between a confederation and a republic.
8. Illustrations and examples
 - a. Give two illustrations of the influence of the Federal acts of reconstruction on Southern political life between 1867 and 1900.
 - b. Name three examples of the action of oxygen.
9. Application of rules or principles
 - a. Would a piece of iron 6 feet long be longer or shorter when heated? Why?
 - b. Would an ordinary pump lift water higher or lower on a mountain than on a plain? Why?
10. Discussion
 - a. Discuss the influence of weather on rocks and soils.
 - b. Discuss the meaning of a climax in a play as to (1) its general nature, (2) its position in the usual play.

IMPROVING THE SCORING OF ESSAY-TYPE QUESTIONS

After the teacher has assured himself that (1) the questions reflect the presence of the complex mental processes which are the objectives of his course and (2) the questions are carefully and accurately made, his next problem is to improve the reliability of scoring such questions. Two procedures will be described, both of which require that the acceptable answers be set down and considered before the scoring begins.

The Sorting or Rating Method

In Sims's preliminary investigations, results obtained from scoring separate questions, one at a time, and adding up the scores from the separate tests were compared with results obtained from rating examinations for general merit. Sims¹ concluded that, of the two, rating for general merit was more economical of time and more reliable. His procedure is roughly described by the following imperatives:

1. After a quick reading, sort the papers into five groups: (a) very superior, (b) superior, (c) average, (d) inferior, and (e) very inferior. The number in each pile is somewhat controlled by the percentages allocated to each pile. The highest and lowest piles are to receive 10 per cent each; the next piles, as we move toward the center, 20 per cent each; and the middle pile 40 per cent. Thus we have about 10, 20, 40, 20, and 10 per cent in the five piles. There was no inclination to use exact percentages.

2. Do not give separate grades to individual questions, but place each paper in its appropriate pile according to its general total merit.

3. Reread the papers, then shift a paper to another pile when such a procedure seems warranted.

4. Give all the papers in the highest pile, A; the second highest, B; and so on until all the papers on the lowest pile receive E.

A similar procedure has been recommended by Rinsland² but the percentages approached more nearly those of the normal curve: 6, 22, 44, 22, and 6 per cent approximately. He advised the raters to "think only of quality in terms of subject matter." Better results are achieved if the names are written on the papers where the rater cannot see them. The reliabilities achieved by correlating two teachers' ratings of the same papers ranged from .67 to .79, with an average of .72.

¹ Sims, Verner, "The Objectivity, Reliability, and Validity of an Essay Examination Graded by Rating," *Journal of Educational Research* (1931) 24:216-223.

² Rinsland, *op. cit.* p. 253.

The Point-score Method of Scoring Essay-type Questions

In the point-score method an analysis is made of the acceptable answers to the questions. It is decided what each part shall receive. In the procedure used in the College Entrance Examination Board, the readers have gotten together and—after consultation with one another—have decided upon the number of points to be attributed to each acceptable item.¹ The result is much more exact because the questions have been carefully constructed and even tried out in a preliminary way on a small group. In grading such a subject as English the reader with the number of items already agreed upon for each part of a question before him scores only a small part of the total examination and records his results on a detached scoring sheet. When the first reader has finished with a paper a second reader proceeds with its grading until many readers have scored the same question. For example, in scoring a question which asks for the recognition and interpretation of metaphors, one point might be given for indicating the metaphors, another for setting forth their meaning, a third for the acceptable comprehension of the passage as a whole, a fourth for becoming aware of the humorous purpose of the passage, and a fifth for composition provided it did not contain jarring errors of grammar. Under such rigorous conditions of construction and scoring, truly remarkable agreements between readers can be had. Thus the reported reliabilities for certain 1937 examinations of the College Entrance Board, obtained by an independent rereading of a sample of the papers, are:²

| Subject | N | Reliability |
|---------------------|-------|-------------|
| Biology..... | 144 | .96 |
| English..... | 1,149 | .84 |
| History, A..... | 49 | .98 |
| Mathematics, A..... | 296 | .97 |
| Spanish..... | 25 | .97 |

Such high reliabilities are most certainly not to be expected under ordinary circumstances, and they give a false impression of the accuracy of the essay-type question. They do show that, with care in construction and agreed-upon scores for analyzed elements, reliability can be

¹ Noyes, E. S., "Recent Trends of the Comprehensive Examination in English," *Educational Record*, Supplement No. 13 (1940) 21:107-119.

² Stalnaker, John M., "Essay Examinations Reliably Read," *School and Society* (1937) 46:671-672.

greatly increased. These more exact procedures in scoring have aided us to preserve the important characteristics of the essay-type question.

In conclusion, it is abundantly clear that both short-answer and essay-type questions are necessary to measure adequately the objectives of the ordinary course. Very clearly have the advantages of the short answer been stated by A. C. Krey, a student who is not himself an expert in test construction:¹

It [the new type of test] is the most efficient device for detecting the student's possession of those separate material elements which, though not the end of instruction, are an essential preliminary to those ends comparable to the shoring which the engineer employs in shaping buildings made of concrete. No other testing device covers so great a range of information in so short a time, or can be graded so quickly and accurately. It may also be used to discover the student's knowledge of the simpler and limited relationship of this material. It may be used to some extent, also, in testing students' ability to apply ideas to new materials, and his possession of the skills involved in the subject. The more advanced and complex stages of these values, however, must as yet be discovered by other forms of test.

While many testers would deny the strictures placed upon the short-answer test for measuring the more complex stages of understanding, they would all agree that the short-answer test covers the largest area in the shortest time.

When comparisons are to be made, contrasts to be indicated, assumptions stated, materials to be summarized or outlined, and deductions made or inferences drawn from a large amount of material, the essay type of test is more efficient and should be used.

SUMMARY

Short-answer tests are an attempt to evaluate more precisely and more completely the results of instruction. They depend for their usefulness upon a more exact definition of the objectives of instruction. Their main strength lies in two major areas. In the first place, they are strong because they can sample in the time available a much larger number of defined outcomes than can the essay type of examination. In the second place, if carefully constructed, they can measure more reliably these very outcomes. They are weakest in the evaluation of the higher mental processes such as judgment and reasoning. These short-

¹ Kelley, Truman L., and A. C. Krey, *Tests and Measurements in the Social Sciences*, p. 482. New York: Charles Scribner's Sons, 1934. By permission.

answer instruments may be divided into two classes; those based on recall and those based on recognition.

Two types of short-answer tests based on recall are presented: the simple-recall type and the completion type. In simple recall the larger unit of instruction is broken down into smaller ones and definite questions are asked that can be answered in a word or in a phrase. Great care must be taken to phrase the question in such a restrictive manner that only one answer will be possible. Acceptable answers must be listed for each item before the scoring begins. In the completion test key words are omitted which presumably can be supplied only by those who are steeped in the material being tested. Considerable ingenuity is needed on the part of the test constructor to provide just enough of the sentence to make the thought intelligible and not enough to give away the answer. Each blank should be of the same length and have in it a number in parentheses. These numbered omissions have, for easy scoring, a vertical column of blanks whose numbers correspond to the blanks in the body of the test.

The second type of short-answer questions is based on the principle of recognition. Several answers are supplied, and the subject to get the answer correct must check the right answer. Four forms of this type of question were set forth: multiple-choice, true-false, matching, and tests of higher mental processes. Of these, matching and multiple choice are most alike. They depend for their efficiency upon the plausibility of all answers and the homogeneity of the answers themselves. The multiple choice has the greatest all-round usefulness. In general, about four or five plausible choices are used for each question, from among which the subject tries to choose the correct answer. The matching technique is more compact, since only one list of answers is necessary for a number of questions. The number of answers in this list may not be more than two more than the number of items to be matched. The sets of answers or matches should be homogeneous among themselves. The true-false form is easy to construct and to score. It is handicapped as a form because there are only two choices and a subject has a 50-50 chance of getting an item correct. Correcting for chance forms a difficult problem. Some experts recommend that all the items be attempted and the score used be the number of items correctly marked. Among the tests aimed at testing the higher mental processes those having to do with perceiving relationships in data and the ability to recognize the limitations of data have been most successful. Such tests are based on the principles of reaching the correct conclusion from data and of checking the right principle on which the correct interpretation depended.

The essay type of question may stimulate the student to exercise

his higher mental processes, to state conditions on which an assumption rests, and to develop a sustained exposition of large numbers of ideas. Furthermore, it permits the outlining and summarizing of great areas of information. Because of these strong points the essay type of question needs to be improved. Such improvement may come in the question and in its scoring. The questions may be improved by incorporating in the question the lines of reasoning which are to be developed. Improvement in scoring accrues from deciding upon the answers to questions before the scoring begins, followed by either (1) a rating of the examinations as a whole and dividing them into appropriate piles, or (2) defining rigidly the points to be scored and then summing the points.

QUESTIONS AND EXERCISES

1. Make a point-by-point comparison of the recall type of short-answer tests with the recognition type. Which type seems to you to furnish the better evaluation?

2. Select an area of information with which you are very familiar. Construct 20 true-false items, 20 multiple-choice items, 10 completion items, and 20 short-answer items. Follow closely the principles laid down for the construction of each type. Which type measures better the higher mental processes involved?

3. How should the wrong items be treated in scoring the true-false type? The multiple-choice type? Can you write a general formula for correcting for chance which will apply to all cases involving guessing? Explain the principle involved.

4. Describe the procedures used in evaluating the use of the higher mental processes. Do you think this process of reasoning should be a defined outcome of our education? Why?

5. What advantages might accrue from the emphasis upon evaluation in the learning process?

6. What are the leading difficulties in measurement of outcomes of education arising out of the use of the essay type of test? What essential outcomes are tested by the essay type of examination which are very difficult to test by the short-answer type?

7. Describe the procedures used for improving the construction and scoring of essay-type tests.

8. Distinguish sharply between the situations suitable for (a) the short-answer test, and (b) the essay-type test.

BIBLIOGRAPHY

CRONBACH, L. J.: "An Experimental Comparison of the Multiple True-False and Multiple Choice Tests," *Journal of Educational Psychology* (1941) 32:533-543.

HAWKES, HERBERT E., E. F. LINDQUIST and C. R. MANN: *The Construction and Use of Achievement Examinations*, Part II, pp. 163-442. Boston: Houghton Mifflin Company, 1936.

KELLEY, T. L., and A. C. KREY: *Tests and Measurements in the Social*

Sciences. New York: Charles Scribner's Sons, 1934.

MICHEELS, W. J., and M. RAY KARNES: *Measuring Educational Achievement*. New York: McGraw-Hill Book Company, Inc., 1950.

MONROE, WALTER S., and RALPH E. CARTER: *The Use of Different Types of Thought Questions in Secondary Schools and Their Relative Difficulty for Students*, Bureau of Educational Research Bulletin No. 14, University of Illinois, 1932.

NOYES, E. S.: "Recent Trends of the Comprehensive Examination in English," *Educational Record Supplement* No. 13 (1940) 21:107-119.

ORLEANS, JACOB S., and GLENN A. SEALY: *Objective Tests*, Chap. XIII, pp. 218-242. Yonkers, N.Y.: World Book Company, 1928.

REMMERS, H. H., and N. L. GAGE: *Educational Measurement and Evaluation*, pp. 146-193. New York: Harper & Brothers, 1943.

RINSLAND, H. D.: *Constructing Tests and Grading*. New York: Prentice-Hall, Inc., 1937.

ROSS, C. C.: *Measurement in Today's Schools*, 2d ed., pp. 103-171. New York: Prentice-Hall, Inc., 1947.

RUCH, G. M.: *The Objective or New-type Examination*, Part II, Chaps. VII-X, pp. 149-280. Chicago: Scott, Foresman & Company, 1929.

SIMS, VERNER: "The Objectivity,

Reliability, and Validity of an Essay-examination Graded by Rating," *Journal of Educational Research* (1931) 24: 216-223.

SMITH, E. R., R. W. TYLER, *et al.*: *Appraising and Recording Student Progress*, Part I. New York: Harper & Brothers, 1942.

STALNAKER, JOHN M.: "Essay Examinations Reliability Read," *School and Society* (1937) 46:671-672.

TRAVERS, ROBERT M. W.: *How to Make Achievement Tests*. New York: The Odyssey Press, Inc., 1950.

TYLER, R. W.: *Constructing Achievement Tests*. Columbus, Ohio: The Ohio State University Press, 1934.

WEIDEMANN, C. C.: "Written Examination Procedures," *Phi Delta Kappan* (1933) 16:78-83.

———: "Review of Essay Test Studies," *Journal of Higher Education* (1941) 12:41-44.

CHAPTER 4

The Testing Program—Achievement-test Batteries

Let us assume that the objectives of instruction of an elementary school have been decided upon and that teacher-made tests have been administered and the results studied. But the outcome is not satisfying, something seems to be lacking. There is no way of deciding for certain whether the pupils are *really* doing as well as schools in other communities. Other questions as to whether the pupils are progressing at the usual rate also arise. Such a condition furnishes a fruitful opportunity for developing a program of testing with standardized tests.

PLANNING FOR THE TESTING PROGRAM

For a program to be most successful, it must have the cooperation of the entire staff. Even a few malcontents can throw a monkey wrench into the machinery. To ensure this desirable cooperation the whole faculty must be involved. The principal, therefore, must call them together and the whole problem of testing must be introduced. It is well in this initial meeting to have someone well versed in testing to present the matter. Suppose now that the faculty votes in favor of such a program. If so, committees are formed to study the areas where testing can be done with the greatest promise of success. After a short while the committees make their reports, thresh out their differences, and define and agree upon their major needs of testing.

DEFINING THE PURPOSE

From the democratic procedures described in the preceding paragraph, suppose that the following purposes emerged:

1. To test pupils in reading for understanding.
2. To determine the level of success of each pupil in each of the subjects of the curriculum according to his age, his grade, and his ability.
3. To study the progress of each pupil in each subject.
4. To discover in which school subject each pupil is strongest and in which, weakest.

SELECTING THE LEADER

It is very important that a responsible leader be in charge of the whole program. Who this will be depends on the circumstances. In small schools the leader is usually the principal or someone appointed by him who has had special training in tests and their interpretation. In larger schools the leader may be a guidance teacher or a member of a bureau of testing. Whoever this leader is, he must have free time for organizing and directing the whole program.

SELECTING THE GRADES AND THE TESTS

Once the purposes of the testing program are clearly defined the selection of tests is made easier. The tests must be selected to carry out the aims and purposes of the program. *Indeed, the direction of the whole program is determined by the defined purposes.* In carrying through our particular program the selection of the grades would be easy, for all grades, except possibly the first, would be included. The selection of the tests would be a more difficult matter.

First and foremost, the tests would be selected to carry out the purposes of the testing program. We must have a good test of reading and a good complete test battery to answer questions about level and progress in the several subjects. An intelligence test is also needed to appraise the level of pupils or classes in relation to their ability.

There are several easy ways to select tests. The easiest way of all is to write to the department of education of your state university and ask for four or five names of tests which cover the areas envisaged in your purposes. You would get a list of suggested tests from which a few tests must be selected to suit the unique situation present in your school. These tests must be ordered and a careful study of them made. This careful study of the tests by the committee develops such an understanding of the test that many possibilities of use, not before thought of, may be discovered. If the leader and an appointed committee of teachers are ready to investigate and decide upon the tests to carry out the program agreed upon, what characteristics of the tests shall they look for?

1. First and foremost, the test must cover the same ground and reflect the same objectives as the instruction in the grades. The content and the methods of test construction in each test must be examined in considerable detail to see that the pupils have had an *opportunity to learn* the answers to the questions which are asked on the test. This curriculum content which was so greatly emphasized in our previous chapter applies here in every particular. In short, the question to be

answered about the tests is: Are the objectives defined and practiced in the grade reflected in the test to be selected?

2. The reliability of the test and its method of determination must be satisfactory. Were the subjects selected from two or three grades or from one grade only? How clear is the manual on this point? (See page 37 of this text.)

3. The instructions intended for the pupils must be clear. Do they contain both illustrations properly filled out and one or two to be filled out by the pupil before the actual testing is begun?

4. The opportunities for interpretation such as the adequacy of the norms such as grade norms, age norms, percentiles, and standard scores must be carefully considered. What other opportunities for individual and class analysis exist?

5. The following practical problems must also be considered: (a) the cost of the tests, (b) the complexity of scoring and the time needed for it, and (c) the time to be allotted to the pupils for taking the tests.

These suggestions, except for the first one, also apply to the selection of intelligence tests. In place of suggestion No. 1, concerned with internal or curricular validity, in intelligence tests we look for external validity. How well does an intelligence test correlate with progress through school, with school marks, and with other tests? In short, how was its validity established and how valid is it for its defined purposes?

Of great aid for selection and evaluation of tests are the *Mental Measurements Yearbooks*, which are under the editorship of Oscar K. Buros.

SELECTING TESTING TIME

There are two seasons of the year in which testing is usually done: fall and spring. There are certain advantages in each of these periods.

If testing is done in the *fall*, it should be done about two weeks after the term begins. Results obtained so early in the term may be used for planning programs of improvement, for grouping of pupils within the class for purposes of instruction, and for deciding upon differential procedures for slow and fast learners. The teacher is not so apt to be so greatly concerned about results that she will coach her pupils in the items of the test. These test scores do not show pupils' standings at the end of the year and in some tests such as arithmetic may reflect the results of forgetting over the summer vacation.

Tests given in the *spring* reflect the results of teaching during the year. Their results are of somewhat greater advantage to the administrators who are interested in how classes and schools stand at the end of

TABLE 2. PLANNING THE TESTING PROGRAM*

TESTING PROGRAM ORGANIZATION CHART

Community: Anytown, U.S.A.

PURPOSES OF THE PROGRAM: 1. To aid teacher in a better understanding of the ability and achievement level of each pupil. 2. To point out subject strengths and weaknesses in each school and in the community.

GRADES TO BE TESTED (Circle):

Intelligence: 1,2,3,4,5,6,7,8,9 Achievement: 1,2,3,4,5,6,7,8,9Other (Give type of test and grades): Metropolitan Readiness—Grade I

MONTH OF TESTING:

Intelligence: September Achievement: October Other: (Readiness) October

TESTS TO BE USED (Indicate name of test, battery, and form):

| INTELLIGENCE | | ACHIEVEMENT | |
|----------------------|------------------------|----------------------|---------------------|
| Pintner-Cunningham | Grade(s) <u>1</u> | Metropolitan—Prim. I | Grade(s) <u>2</u> |
| Pintner-Durost | Grade(s) <u>3</u> | “ —Prim. II | Grade(s) <u>3</u> |
| Pintner Intermediate | Grade(s) <u>6 + 8+</u> | “ —Elem. | Grade(s) <u>4 5</u> |
| | Grade(s) _____ | “ —Inter. | Grade(s) <u>6-7</u> |
| | Grade(s) _____ | “ —Adv. | Grade(s) <u>8</u> |

OTHER TESTS

Metropolitan Readiness Grade(s) I _____ Grade(s) _____DIRECTOR OF THE PROGRAM Miss Mary Drake (Elem. Supervisor)EXAMINER(s): Psychologist _____ Principals _____ Teachers X Others _____SCORER(s): Teachers(Individual) _____ Teachers(Group) X Clerks _____ Machines _____

Others _____

CHECK SCORER(s): Teachers(Individual) _____ Teachers(Group) X Clerks _____

Others _____

METHOD OF TEST DISTRIBUTION Tests will be packaged in the principal's office and distributed at the teachers' meeting.

REPORTS TO BE MADE:

BY THE TEACHER: Profile Chart X Class Record X Class Analysis Chart XPermanent record X Other summaries _____BY THE PRINCIPAL: School summary X Other summaries _____BY THE PROGRAM DIRECTOR: Administrative summary X Other summaries _____

TEST RESULTS TO BE RECORDED IN TERMS OF:

INTELLIGENCE: Ratio IQ _____ Deviation IQ X Mental Age _____ACHIEVEMENT: Standard score _____ Grade Equiva. (Trad.) Age Equiv.

Percentiles (Trad.) _____

SCHEDULE OF TEACHER CONFERENCES: Before testing (Date) September 20Before scoring(Date) October 25 For interpreting results(Date) November 15NECESSARY REARRANGEMENTS OF THE DAILY SCHEDULE Assembly period will be omitted on Wednesday, October 27

* From *Planning the Testing Program*, by permission of World Book Company, Yonkers, N.Y.

TABLE 2. TESTING PROGRAM ORGANIZATION CHART (*Continued*)
TESTING SCHEDULE

| Day (date) | Hour | Test | Grade | Adm. time, minutes |
|------------|------|------------------------------|-------|-----------------------|
| Monday | 9 AM | Pintner-Cunningham | 1 | 25 |
| | | Pintner-Durost | 3 | 45 |
| | | Pintner Intermediate | 6 + 8 | 45 |
| Monday | PM | | | |
| Tuesday | 9 AM | MAT Prim. I—Tests 1,2,3 | 2 | 30 (app.) |
| | | MAT Prim. II—Tests 1 + 2 | 3 | 40 (app.) |
| | | MAT Elem.—Tests 1 + 2 | 4 + 5 | 35 |
| | | MAT Inter.—Tests 1 + 2 | 6 + 7 | 35 |
| | | MAT Adv.—Tests 1 + 2 | 8 | 35 |
| Tuesday | 2 PM | MAT Prim. I—Test 4 | 2 | 15 (app.) |
| | | MAT Prim. II—Tests 3 + 4 | 3 | 30 (app.) |
| | | MAT Elem.—Tests 3 + 4 | 4 + 5 | 65 |
| | | MAT Inter.—Tests 3 + 4 | 6 + 7 | 80 |
| | | MAT Adv.—Tests 3 + 4 | 8 | 80 |
| Wednesday | 9 AM | MAT Prim. II—Test 5 | 3 | 15 (app.) |
| | | MAT Elem.—Tests 5 + 6 | 4 + 5 | 35 (app.) |
| | | MAT Inter.—Tests 5 + 6 | 6 + 7 | 40 |
| | | MAT Adv.—Tests 5 + 6 | 8 | 50 |
| Wednesday | 2 PM | MAT Inter.—Tests 7,8,9, + 10 | 6 + 7 | 60 |
| | | MAT Adv.—Tests 7,8,9, + 10 | 8 | 60 |

the year. Teachers are more likely to teach the particular items present in the test, which spoils the test results, since the items are representative samples of a much larger number. The test results may also be used for grouping pupils in class the next fall, though in this respect they are not of the greatest use because of differential forgetting during the summer vacation. The author leans toward autumn because he is most interested in the use of tests for instructional purposes.

Let us suppose now that the season of testing has been decided upon. There still remains the class scheduling which must be arranged in such detail that every teacher will know *exactly* when the tests are to be given.

The Testing Program Organization Chart, Table 2, contains complete details. It lists the grades to be tested, time, director of the program, lists of tests, etc., and the testing schedule. The time for administering each part of each test is an essential and important detail for complete planning. That part of the chart which gives the day, the time of day, the names of the tests, the grades, and the amount of time required for administering each test is of especial interest in the present connec-

tion. Some such detailed schedule should be formulated before the testing is begun.

ADMINISTERING THE TESTS

The teacher is the one who must administer the tests. In some programs planned for special purposes a member of a trained staff of testers may administer the tests. For the ordinary testing program, designed to understand more accurately the achievement and intellectual abilities of pupils so that improvement in instruction may be facilitated, the teacher gives the tests.

To do this job well the teacher must divorce himself from his role as teacher and assume a new one, that of tester. To do this well he must become thoroughly acquainted with the tests to be used. One of the best ways to do this is to go through the entire procedure: (1) read the instructions, (2) take the test, (3) score the test, and (4) interpret it. In the first place *know the instructions* so well that most of the testing time may be used in watching the subjects. Wise testers go through the manual and mark in red (1) what has to be read aloud, (2) where the instructions begin, and (3) above all, where the timing is located. If samples of the test are given to be worked out, he must see that they are done properly. The teacher's job as tester is to see that the pupils (1) understand the directions, (2) do not cheat, (3) work continually and faithfully, and (4) have a quiet place to work without interruptions. To avoid interruptions place a placard on the door of the classroom: "Testing Going On—Do Not Disturb."

Timing is most important. Secure a stop watch, if possible; if not, a watch *with a second hand*. *Write down the time the test begins and when it ends*. While the tester in no way hints or suggests what the answer to an item is, neither is a good tester a "deadpan." The good tester encourages children to do their best, keeps their attention on their work by asking them to try for a good record, and in every way encourages them to do their best. The teacher, of course, *must not aid the pupils* in answering the items *either by direct aid or by suggestion*. It is necessary for the leader to call a meeting of the faculty and go through the test's details point by point as has been suggested in the preceding paragraph. He must emphasize, as must we all, that *instructions and directions must be followed exactly*. Unless the instructions are carried out meticulously, comparisons with norms and with records of other grades cannot be usefully made.

SCORING THE TESTS

For best results the teachers are once again called together by the leader and the details of scoring carefully reviewed. All standardized

tests give detailed instructions for scoring. In some large municipalities the pupils write their answers on a separate sheet and the scoring is done by an International Business Machine, commonly called IBM. But in our plan the teacher scores the papers. It is always an onerous task and takes several hours of work. Many devices have been developed for shortening the scoring time, such as window stencils, stiff cardboard with lists of answers, and squares in which the right answer enters a cross.

The experience of the author recommends the following procedure for its speed and enjoyment. If there are eight subtests, obtain the services of nine teachers who sit around a large table. Each teacher becomes responsible for scoring a single test. Teacher No. 1 scores the first test and folds the paper back to Test 2, which the second teacher scores; he then turns to test 3, which teacher No. 3 scores, etc. The ninth teacher brings the scores forward to the front of the test, enters them in the proper place and adds them up. Once this procedure has been started, the scored tests roll off the line in a continuous stream. After a little practice each teacher practically memorizes the answers for his test and the work moves rapidly.

For accurate scoring it is necessary for the scoring to be checked by a person not concerned in the first scoring. Samplings of about one test in five for checking are adequate. Errors most likely to occur are concerned with correct adding, computing averages or medians, and scoring those items which are scored by the right-minus-wrong ($R - W$) technique.

INTERPRETING AND UTILIZING THE RESULTS

Proper interpretation and utilization of results are likely to be the weakest links in the chain of testing. But without them the whole testing program is without value for improving the processes and materials of education. The problem of interpretation hinges on the arrangement of scores in such a manner that their meaning is immediately apparent. Generally speaking, some sort of derived scores are more meaningful than the raw scores secured from the tests. Samples of derived scores are age and grade scores, I.Q.s, and percentiles.

The first illustration of interpreting scores will be the record of a single child such as would appear with slight variations in any completed record. Craig Smith in Fig. 2 is in grade 7.2, or two-tenths of the distance through the seventh grade. By looking at Average Achievement at the bottom of the table, you note that his score is 7.5. His achievement is somewhat *above* his grade standing. To interpret this score more accurately we would need his I.Q. also. If his I.Q. were 100 this would be good; if it were 125, he would not have fully achieved up

METROPOLITAN ACHIEVEMENT TESTS.

ADVANCED BATTERY — COMPLETE: FORM 5

Name Craig Smith Boy 1 Girl ...
 Teacher Mrs. White Grade 7.2 School P.S. #3
 City Wintertown County Jones State Arizona

| Test | STANDARD SCORE | GRADE EQUIVALENT | Percentile |
|--|----------------|------------------|------------|
| 1. Reading | 215 | 8.2 | 65 |
| 2. Vocabulary | 215 | 8.0 | 65 |
| Average Reading | \times | (8.1) | |
| 3. Arithmetic Fundamentals | 178 | 5.4 | 5+ |
| 4. Arithmetic Problems | 194 | 6.0 | 25- |
| Average Arithmetic | \times | (5.7) | |
| 5. English | | | |
| I. Language Usage | | | |
| II. Punct. and Cap. Total (Parts I and II) | 216 | 8.0 | 65+ |
| III. Grammar Total (Parts I, II, and III) | 225 | 8.5 | 75+ |
| 6. Literature | 214 | 7.9 | 65- |
| 7. Social Studies: Hist. | 228 | 10.1 | 85- |
| 8. Social Studies: Geog. | 214 | 8.1 | 65- |
| Average Social Studies | \times | (9.1) | |
| 9. Science | 202 | 6.6 | 40- |
| 10. Spelling | 193 | 5.9 | 20- |
| Average Achievement | \times | 7.5 | |

*Do not include when figuring average achievement.

1946 11 20
 Year Month Day
 Date of Testing

1933 10 19
 Year Month Day
 Date of Birth

Age 13 yrs. 1 mos.

FIG. 2. Completed title page for the advanced battery. (By permission of World Book Company, Yonkers, N.Y., 1947.)

TABLE 3. CLASS
(Metropolitan)

| Name | C.A. | M.A. | I.Q. | 1 Read. | 2 Vocab. | Ave. Read. | 3 Arith. Fund. | 4 Arith. Prob. |
|----------------------|-------|-------|-------|------------|-------------|---------------|----------------------|----------------------|
| 1. Abrams, John..... | 11-6 | 12-8 | 110 | 10.1 | 9.4 | 9.8 | 6.7 | 6.9 |
| 2. Boyd, Sue..... | 10-10 | 11-4 | 105 | 6.1 | 7.7 | 6.9 | 6.0 | 6.3 |
| 3. Cady, Arthur..... | 12-3 | 10-8 | 87 | 5.9 | 6.5 | 6.2 | 5.9 | 6.0 |
| | | | | | | | | |
| 25. Waters, Roy..... | 12-4 | 11-6 | 93 | 5.6 | 5.8 | 5.7 | 5.2 | 5.6 |
| Median of 25..... | 12-2 | 12-6 | 103 | 6.2 | 6.8 | 6.5 | 5.8 | 5.8 |

to the level of his ability. By again examining the column called Grade Equivalent we find a variation from 5.4 in arithmetic fundamentals to 10.1 in social studies. Here is a variation of over 4.5 grades. Craig undoubtedly needs special work in arithmetic. The first thing to do would be to go over his arithmetic tests with him to discover in what processes he had made his errors and to enlist his cooperation in planning for his improvement.

A second way to study the results of testing is by means of the usual class record sheet. In one which is before the author, there are 25 names arranged alphabetically like a class roll. For each child there is arranged along the top a record of C.A., M.A., and I.Q. and grade equivalents in 10 different subjects taught in the elementary school: reading, arithmetic, English, etc. This furnishes three items additional to those of Craig Smith in Fig. 2. These are C.A., M.A., and I.Q. (Table 3). One can now study the pupils' grade equivalents in the light of their chronological age and ability as measured by an intelligence test. In this table, one child, 11 years and 6 months of age, has a grade equivalent of 10.1 in reading; while another child of 12 years and 3 months scores 5.9 in the same subject. In the first child it will be noticed that his M.A. is 12-8 and his I.Q. is 110. In the case of the second child, his M.A. is only 10-8. It is thus seen that his reading score is more closely related to the M.A. than the C.A. From such a table also the grade equivalent of each member of the entire class in reading or in any other subject may be inspected. If desired, the mean of the class in reading may be computed. In this sixth-grade class, grade equivalents in reading vary from 10.6, the highest, to 3.8, the lowest, with a median of 6.1. You can readily see that the individual with the grade equivalent of 3.8 has a hard time trying to read sixth-grade materials.

RECORD SHEET*

Battery, Grade 6)

| Ave. Arith. | 5 Eng. | 6 Lit. | 7 Hist. | 8 Geog. | Ave. Soc. Stds. | 9 Science | 10 Spell. | Ave. Ach't |
|----------------|-----------|-----------|------------|------------|-----------------------|--------------|--------------|---------------|
| 6.8 | 10.6 | 10.5 | 10.9 | 11.3 | 11.1 | 9.2 | 7.9 | 9.4 |
| 6.2 | 7.5 | 9.3 | 5.3 | 5.9 | 5.6 | 6.3 | 7.7 | 6.8 |
| 6.0 | 7.5 | 6.8 | 6.9 | 3.7 | 5.3 | 6.3 | 8.3 | 6.4 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 5.4 | 6.1 | 5.2 | 5.5 | 5.7 | 5.6 | 5.7 | 6.1 | 5.6 |
| 5.8 | 7.0 | 6.8 | 5.7 | 5.5 | 5.6 | 6.3 | 7.1 | 6.1 |

* By permission of World Book Company, Yonkers, N.Y.

A third method of understanding quickly what these scores mean uses a graphical representation by means of which a child's score may be compared with the average of *his* class. In Fig. 3 appears such an arrangement. In the first place you will note that the pupil's I.Q. is 87 while that of the class is 96.5. This low I.Q. explains somewhat his low score on arithmetic problems and reading for understanding. The greatest retardation of this child in comparison with the class comes in geography and history followed closely by vocabulary and literature. The class profile shows a class low in arithmetic, good in English, and poor in geography.

The fourth and final illustration of the interpretation of scores appears in Fig. 4, a normal progress chart. In this chart, designed to show "a cumulative attainment record for grades 4-8," the percentile rank is shown with the testing record at the bottom. The graph, itself, shows the weakness of using percentile ranks for purposes of expressing growth. At the very beginning, in the spelling column, the child seems to have grown worse instead of better, but if we examine the grade equivalent in the record at the bottom we find he had improved from 6.7 to 7.0. He had certainly dropped relatively as shown by percentile rank but he had improved absolutely as shown by the grade equivalent. It thus appears that a graph of growth based on grade equivalents would more nearly show at a glance the true facts. One advantage of this graph is the I.Q. score with the possible implication that this child was neither achieving nor progressing up to his ability.

Teachers, supervisors, and administrators use such records as the ones described in a variety of ways. For this reason the records should be carefully filed in the pupil's folder and entered on his cumulative record card. The growth of the pupil from year to year can thus be studied and his total personality more nearly understood.

To return to our immediate problem described at the beginning of this chapter, it is evident that (1) pupils have been tested for their understanding of reading material, (2) the level of each pupil has been determined in each of the subjects of the curriculum and interpreted in the light of the pupil's age, grade, and ability, (3) procedures have been described which portray the progress of children in subjects, and (4) records of tests have been introduced from which the pupil's strongest and weakest subjects could be easily seen.

From such accumulated data, programs of instruction can more easily be fitted to the level of growth attained by each pupil, groupings of children within each subject for purposes of instruction can be more easily made, and the level of achievement attained as compared with national norms more easily understood.

Name Thomas Richards Date October, 1947
Teacher Miss Brown Grade 6.2 School McKinley
City Wintertown County Jones State Arizona

INDIVIDUAL PROFILE CHART

METROPOLITAN ACHIEVEMENT TESTS: INTERMEDIATE BATTERY—COMPLETE

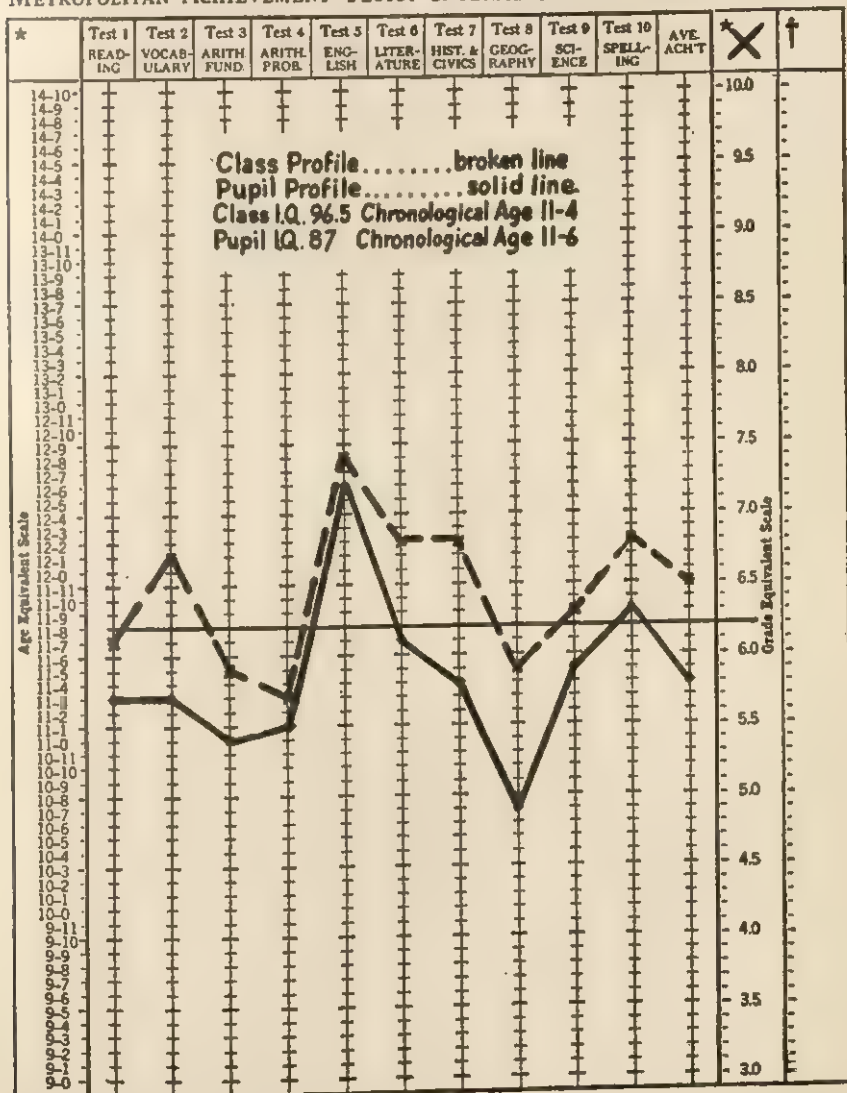
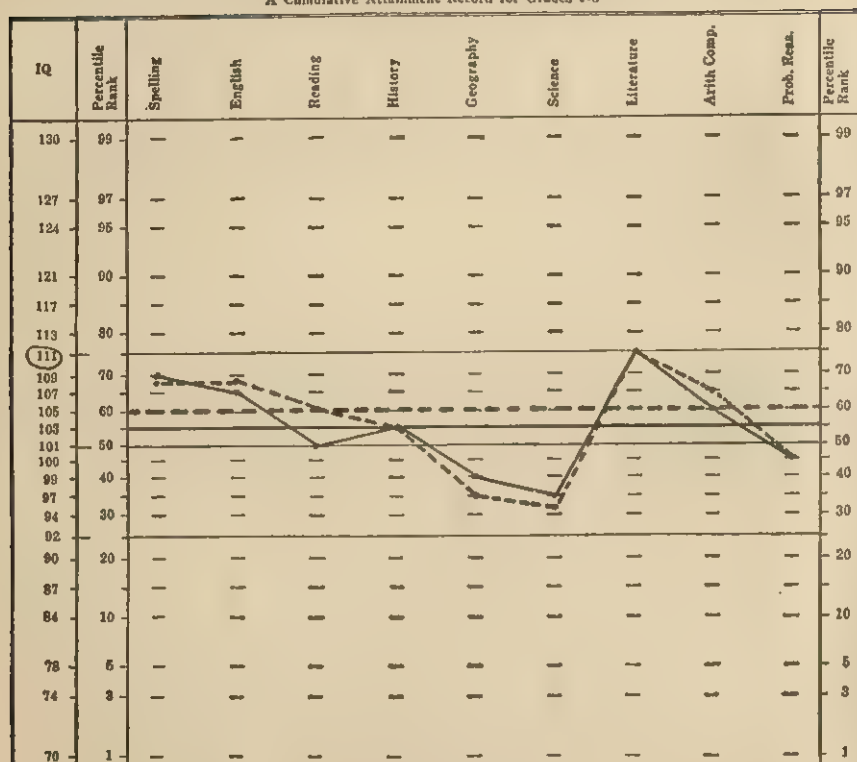


FIG. 3. Completed profile chart (intermediate battery), comparing the performance of an individual with that of the class in terms of traditional grade equivalents. (By permission of World Book Company.)

COORDINATED SCALES OF ATTAINMENT

Normal Progress Chart

A Cumulative Attainment Record for Grades 4-8



TESTING RECORD

| Intelligence | | | 1st Testing | | | 2nd Testing | | | 3rd Testing | | | 4th Testing | | | 5th Testing | | |
|---------------|---------|------|-------------|-----|----------------------------------|-------------|-------|------|-------------|-----|-------|-------------|---------|-----|-------------|------|---------|
| Test and Form | Date | MA | CA | IQ | Coordinated Scales of Attainment | Key | Score | P.R. | Gr. Eq. | Key | Score | P.R. | Gr. Eq. | Key | Score | P.R. | Gr. Eq. |
| Kuhlm.-And. | 1/20/47 | 11-9 | 10-6 | 111 | Spelling | 57 | 70 | 6.7 | 53 | 68 | 7.0 | | | | | | |
| | | | | | English | 55 | 65 | 6.4 | 53 | 68 | 7.3 | | | | | | |
| | | | | | Reading | 51 | 50 | 5.7 | 50 | 60 | 6.5 | | | | | | |
| | | | | | History | 52 | 55 | 5.9 | 48 | 55 | 6.2 | | | | | | |
| | | | | | Geography | 47 | 40 | 5.0 | 43 | 35 | 5.4 | | | | | | |
| | | | | | Science | 46 | 35 | 4.9 | 42 | 32 | 5.2 | | | | | | |
| | | | | | Literature | 55 | 75 | 6.8 | 53 | 75 | 7.3 | | | | | | |
| | | | | | Arith. Comp. | 54 | 60 | 5.9 | 50 | 65 | 6.5 | | | | | | |
| | | | | | Prob. Res. | 49 | 45 | 5.4 | 47 | 45 | 5.8 | | | | | | |
| | | | | | Median | 52 | 53 | 5.9 | 50 | 60 | 6.5 | | | | | | |

Published by EDUCATIONAL TEST BUREAU, Educational Publishers, Inc., Minneapolis - Nashville - Philadelphia

FIG. 4. Progress chart of individual (percentiles). (By permission.)

Programs of achievement testing usually begin with test batteries which survey the various areas taught in the elementary school. If the records from such tests show that the average grade-equivalent scores in some area, say arithmetic, are much lower than desirable, then an achievement-test battery consisting only of tests of arithmetic is used. Survey tests, then, furnish the general level of achievement together with analysis of the total into levels of various subjects. The separate achievement test of a single school subject covers all areas of this subject in greater detail and provides, for this reason, greater opportunities for analysis and diagnosis. The diagnostic test constructed after careful investigations of errors and misunderstandings offers still greater opportunities for analysis and diagnoses of those errors and misunderstandings which retard so greatly the learning process.

Generally speaking, then, testing proceeds as follows: (1) achievement-test batteries, (2) subject-test battery, and (3) diagnostic tests. For this reason, it seems logical to present discussions in this order. Our first treatment of standardized tests will be, therefore, of achievement-test batteries.

DEVELOPMENT OF ACHIEVEMENT-TEST BATTERIES

In the early days of testing achievement, there was an attempt to narrow the function measured, so that its measurement could be more exact. Thus among the scientific tests first developed were Stone's Reasoning Test in Arithmetic, Thorndike's Handwriting Scale, Curtis's Arithmetic Tests, and Buckingham's Spelling Scale. It is true that there were some scales involving more complex processes such as the Thorndike-McCall Reading Scale, the Hillegas Scale for Measuring Composition, and the Hotz Algebra Scales. But the tendency was toward simplifying or abstracting the function so that it could be measured accurately. In all cases, there was no attempt to measure several areas with one test.

Certain practical difficulties developed as a result of this procedure. In the first place, it was expensive and time-consuming to administer five or six tests at different times. Furthermore, even if the tests were administered and properly scored there was no way of comparing the results, say, of two tests. For example, suppose one test such as vocabulary contained 40 items and another such as arithmetic reasoning had only 15 items. It is clear that a score of 10 on one would not be comparable to a score of 10 on the other. Another weakness which characterized these earlier tests was that they rarely if ever had more than two forms. Studies of children's growth require more than two forms for the third or fourth testing. Many of our test batteries today

have four or five equivalent forms which may be used to study children's educational growth over a period of several years.

Two principles of construction also characterized these earlier tests. In the one, all the items were of equal difficulty. The score was the number of items finished in a defined time. Not *how hard* but *how many* was the question to be answered. Examples of tests constructed on this principle are the Courtis Tests of Arithmetic and the Ayres Tests of Reading. The quality of the work was controlled by counting only the problems that were correctly done. In the second method, the items of the test increased in difficulty from the first to the last one. Time was controlled by allowing for the test ample time for all to finish. The attempt was made to have some problems easy enough for all to make some score and difficult enough so that none would finish. If there were several subjects who (1) made no score, or (2) finished all the items in the test, their scores were said to be *undistributed*. In some of the tests, such as the Woody Arithmetic Test, the items were carefully scaled in difficulty by making use of the number of correct answers which an item received. An item on which the subjects scored 90 per cent correct was an easy one, while an item with only 5 per cent of the scores correct was a difficult one. Today the great majority of test items are constructed according to method No. 2, *i.e.*, they *increase in difficulty* within the test.

It was fortunate for the testing movement that three highly competent individuals pooled their resources to develop the first comprehensive test battery. These three men were L. M. Terman, Giles M. Ruch, and Truman Lee Kelley, and the test was called the Stanford Achievement Test. There were other attempts to combine tests before this one, but none of them achieved the completeness or exerted the influence of the Stanford Achievement Test. Since the publication of these tests, or this set of tests, many test batteries have been constructed, but all of them have certain characteristics in common with this first achievement-test battery.

Here are some of these common characteristics. In all of them several subject areas are used and all tests are standardized on the same population. It is then possible to make direct comparisons between the progress of pupils in one subject and that in another or, as in later scales, several others. One could thus say with some degree of assurance that John's reading ability was definitely above his ability in the fundamentals of arithmetic. As these instruments developed two problems were continually being met. The first of these arose because of the factor of age. One fourth grade might score about the same as another, but the children of one might be a year older than those of the other. It became apparent that by withholding promotions any

grade could automatically be brought to the desired level in attainment. Thus the problem of age needed to be definitely taken into account. The second problem of prime importance came in attempting to make comparisons between children on tests in which the items composing each test were significantly different in *number*. A reading test might have 50 items while an arithmetic problems test might have 10 items. A score of 10 on these two tests would have a quite different meaning. Two procedures are most commonly used to meet this problem. One of them is the use of the T-score, or standard score; the second is the grade position. Many test batteries use both these techniques.

The T-score really grew out of the standard score. As experience grew with scores from large populations of children it was apparent that many of them grouped themselves near the mean but that fewer and fewer scores occurred as one went further and further out from the mean in either direction. In short, the normal curve fitted closely enough the scores thus arrived at. It was also recognized that the standard deviation was (see page 507 for the statistics involved) the best indicator of dispersion or deviation from the mean. By putting these two ideas together there was developed the T-score with a mean of 50 and a standard deviation of 10. In the original computation as developed by McCall, 5 standard deviation units were used in either direction from the mean. This would mean a continuum beginning at 0 and going to 100. Progress of pupils could thus be measured in terms of standard units which are as nearly equal to each other as any unit of measurement thus far discovered in education. It became apparent as time went on that it was not necessary to have 50 as a mean and a standard deviation of 10. One might use a mean of 100 or 150 and a standard deviation of 20 with equal accuracy among the units. Semi-interquartile or Q units have also proved useful. At any rate, raw scores are transmuted into these T-scores, direct comparisons are made between the several tests, and profiles are drawn from them to aid the eye in comprehending immediately the total pattern of the child's development. Note especially that this condition holds only *if all the tests are standardized on the same population*.

The second procedure used to equate scores from tests of different numerical length is the grade equivalent. The grade equivalent has the advantage of being easily understood. A score of 10 on the arithmetic problem solving test might be accomplished by the average child in grade 4 while a score of 10 on a reading test might be attained by the average of grade 3.

These two scores could now be transmuted into grade equivalents as follows:

| | Score | Grade Equivalent |
|---------------------------|-------|------------------|
| Arithmetic reasoning..... | 10 | 4.0 |
| Reading..... | 10 | 3.0 |

If, therefore, a child had a score of 10 in each of these two subjects, we can then say that there is a difference of one whole grade between his ability in one subject and his ability in the other. The grade unit, while very practical, is not as accurate as the T-score; *i.e.*, the growth of a grade at one level of advancement does not equal the growth of a grade at another level. While grade equivalents must be used with caution they are high in practicality.

COMPLETE BATTERIES

For purposes of study test batteries may be divided into two groups. The first of these attempts to sample nearly all the outcomes of the elementary school. Not only are reading, arithmetic, spelling, and language included but also literature, the social sciences, and elementary science. Such batteries are long enough to require four sittings of the children who otherwise might tire of such a long examination. In general, large pools of items are selected from textbooks and courses of study and are then submitted to experts in the several areas for critical evaluation. From this pool are selected items which are arranged under different forms of the tests. Preliminary tryouts are made and the final form of the test determined. Norms are then established by administering the tests to hundreds of thousands of children, in some cases 300,000 or more, distributed throughout the country, thus establishing national norms. Illustrations of this type of test are (1) the Stanford Achievement Test, and (2) the Metropolitan Achievement Tests.

The Metropolitan Achievement Tests and the Stanford Achievement Test are alike in many respects. Each of them covers the greater part of the elementary school curriculum. They both have tests ranging from grade 2 through grade 8, and both of them have batteries extending over a few grades rather than all the grades. For example, the batteries of the Metropolitan and of the Stanford Achievement Test are as follows:

| Metropolitan | Stanford Achievement |
|---|------------------------------------|
| Primary I—grade 1 | Primary—end of grade 2 and grade 3 |
| Primary II—grade 2 | Intermediate—grades 4-6 |
| Elementary—grades 3 and 4 | Advanced—grades 7-9 |
| Intermediate—grades 5 and 6 | |
| Advanced—grades 7 and 8 and first half of grade 9 | |

Two differences appear from the outline. The Metropolitan includes tests for grades 1 and 2 separately, and they have more batteries. The advantage in having a test cover fewer grades lies in the fact that many facts have been taught in the upper grades which the children in the lower grades have not learned. These unknown problems tend to discourage some students and make them feel that the test is unfair.

CONTENTS OF THE TWO TESTS

| Metropolitan (Advanced Battery) | Stanford Achievement (Advanced Battery) |
|------------------------------------|--|
| 1. Reading | 1. Paragraph meaning |
| 2. Vocabulary | 2. Word meaning |
| 3. Arithmetic fundamentals | 3. Language usage |
| 4. Arithmetic problems | 4. Arithmetic reasoning |
| 5. English | 5. Arithmetic computation |
| 6. Literature | 6. Literature |
| 7. Social studies: history | 7. Social studies I (history) |
| 8. Social studies: geography | 8. Social studies II (geography) |
| 9. Science | 9. Elementary science |
| 10. Spelling | 10. Spelling |

In general, the content of these tests is much alike. The specific content of the tests differs widely, of course. Both are standardized on subjects located in widely different places, and both have high reliability. Profiles can be easily drawn from the results of each test and grade and age placements read from tables. It might be concluded also that both have the same weak points. Neither has made specific provisions for diagnosing causes of low standing in any area and both of them lean heavily on factual information. In some cases small facts are lifted bodily from their associations and made into a test. The names of books and their authors, what the Vikings called their stories, where the Po Valley is - these are samples taken at random from the Stanford Achievement Test. From the Metropolitan, samples are who Orpheus was, what Arachne was skilled in, and who the first settlers of Saint Augustine were. Further general discussion about the strength and weaknesses of these tests will appear at the end of this section.

TEST BATTERIES OF FUNDAMENTALS

The other type of test concentrates on what might be called the fundamentals which must be learned whether one is a conservative or a progressive in his educational philosophy. The constructors of these tests are skeptical about objective tests in literature and social science. Many of them fear that the factual content which lends itself so easily to test construction does not represent the best outcomes of instruction in these fields. They argue that those areas where hierarchy

of habits prevail, as in language or arithmetic, can be most satisfactorily tested. In the second place, these constructors might say that in spite of the great length of such batteries as the Metropolitan it is impossible to arrange techniques for satisfactory analysis or diagnosis of the test scores. They, therefore, would limit their testing to reading, language usage, arithmetic, spelling, study techniques and, in one case, handwriting. In this type of test special arrangements are made for diagnosis and analysis of each area tested. Illustration of this type are (1) the Iowa Every-pupil Tests of Basic Skills, and (2) the California Achievement Tests.

IOWA EVERY-PUPIL TESTS OF BASIC SKILLS

| Test | Time, minutes | |
|--|---------------|----------|
| | Elementary | Advanced |
| A. Silent reading comprehension..... | 46 | 68 |
| I. Reading comprehension..... | 36 | 58 |
| II. Vocabulary..... | 10 | 10 |
| B. Work-study skills..... | 47 | 77 |
| I. Map reading..... | 11 | 28 |
| II. Use of basic references..... | 8 | 5 |
| III. Use of index..... | 8 | 10 |
| IV. Use of dictionary..... | 12 | 17 |
| V. Alphabetization..... | 8 | 17 |
| C. Basic language skills..... | 46 | 55 |
| I. Punctuation..... | 11 | 4 |
| II. Capitalization..... | 8 | 11 |
| III. Usage..... | 13 | 18 |
| IV. Spelling..... | 8 | 12 |
| V. Sentence sense..... | 6 | |
| D. Basic arithmetic skills..... | 57 | 63 |
| I. Vocabulary and fundamental knowledge..... | 12 | 15 |
| II. Fundamental operations, whole numbers, fractions and decimals..... | 20 | 30 |
| III. Problems..... | 25 | 18 |

The Iowa Every-pupil Tests of Basic Skills are composed of two batteries, with four parts to each battery, as outlined in the accompanying table. The contents shown for each part are those of the elementary battery. The advanced battery contains more complicated

material in the same areas, with three exceptions: (1) instead of alphabetization in Test B it substitutes reading graphs, charts, and tables; (2) in Test C it omits sentence sense, and (3) in Test D it adds to Part II, percentage.

From the contemplation of this table and from the study of the test itself it is clear that by sacrificing breadth this test has achieved depth. Its test of work-study skills is very complete and may aid greatly in locating strong and weak points. A real aid in locating difficulties occurs in the test of the vocabulary and in the test of the fundamentals of arithmetic.

Let us consider the reading test. In the advanced battery there are only four sections to be read, but each section is made up of four or five paragraphs and covers a large page. There is room here for a unit of thought to be developed and an opportunity to ask questions involving both the content of the paragraph and the interrelations between paragraphs. Further study of this test appears in connection with our treatment of social studies.

The California Achievement Tests limit themselves pretty largely to the same area of testing as the Iowa basic-skills test but organize their material more nearly like that of the more complete batteries.

This test provides also a handwriting scale by means of which the handwriting of the words spelled may be rated. On the back of the flyleaf in each pupil's test paper there is a device to record the percentages of errors in the various sections of the test. The page numbers on which the opportunities for these errors occurred are written in. Table 4 on pages 88 and 89 is a sample.

Some of the procedures used for testing are different from those of other tests. For example, in the reading test of the elementary battery the first part of the test has to do with word forms. Are the two words "same" or "different"? Not only do the words increase as to length and complexity, but the printing varies from ordinary printing to the use of capitals and italics—one word in capitals and the other in italics, etc. Vocabulary is frequently presented with the words' opposites as well as with their similars. The test on following directions resembles an intelligence test. There are from 18 to 21 different parts. It is on these parts that the diagnosis of errors is based. Perhaps these short parts on which inferences are based constitute the weakest characteristics of the tests. For example, the table of contents test has only six topics and the test for using the index only six items. Punctuation is tested with only four sentences which are to be properly punctuated. On the other hand, the test for the middle grades are overloaded with arithmetic fundamentals, of which there are four large pages and 80 examples. Worst of all, perhaps, is an English-usage test based on

CALIFORNIA ACHIEVEMENT TESTS

| Primary battery, grades 1-3 | | Elementary battery, grades 4-6 | | Intermediate battery, grades 7-9 | | Advanced battery, high school and college | |
|--|------------------|--|---------|---|------------------|--|------------------|
| Contents | Time | Contents | Time | Contents | Time | Contents | Time |
| 1. Reading vocabulary A. Word form B. Word recognition C. Meaning of opposites D. Following directions E. Directly stated facts F. Interpretations | 14 min. | Same types D. Meaning of similarities | 12 min. | A. Mathematics B. Science C. Social science D. General E. Following directions F. Reference skills G. Interpretations | 12 min. | Same types | 16 min. |
| 2. Reading comprehension D. Following directions E. Directly stated facts F. Interpretations | 20 min. | Same types G. Reference skills | 23 min. | | 38 min. | Literature Same types | 34 min. |
| 3. Arith. reasoning A. Number and sequences B. Money C. Number and time D. Signs and symbols E. Problems | 22 min. | A. Number concept B. Signs and symbols C. Problems | 16 min. | A. Number concepts B. Symbols and rules C. Numbers and equations D. Problems | 30 min. | Same types | 30 min. |
| 4. Arith. fundamentals F. Addition combinations G. Subtraction H. Multiplication I. Problems | 28 min. | Same types Adds division | 44 min. | Same types | 44 min. | Same types | 38 min. |
| 5. Language A. Capitalization B. Punctuation C. Spelling D. Handwriting | 16 min. | Same types Adds words and sentences | 25 min. | A. Capitalization B. Punctuation C. Words and sentences D. Parts of speech E. Spelling F. Handwriting | 26 min. | Same types Adds grammar | 32 min. |
| Total | 1 hr. 40 min. | | 2 hr. | | 2 hr. 30 min. | | 2 hr. 30 min. |

only 10 items which check the difference in usage between such words as "did" and "done," "those" and "them," "seen" and "saw," and "throwed" and "threw."

The evidence points to competent diagnosis in the areas of arithmetic and reading but not in language. Even the inferences concerning weaknesses in reading would be based on rather slim evidence when individual sections are used. The results of analysis would only be tentative and suggestive, with nothing of the finality secured from the test as a whole.

There are other test batteries which follow more or less closely the Stanford and Metropolitan batteries. The Unit Scales of Attainment (recently changed to Coordinated Scales of Attainment) furnish good tests at every level of the elementary school, as do the Gray-Votaw General Achievement Tests. There are also the Modern School Achievement Tests.

Detailed accounts of the tests of school subjects appear under their various headings in this text.

EVALUATION OF TEST BATTERIES

These test batteries furnish very important facts which are of aid in guidance of pupils toward defined objectives and in the appraisal of the results achieved. The results of these tests when carefully given and scored are more accurate and dependable than facts gathered from any other source. Nor are they lacking in comprehensiveness. Indeed, the more elaborate ones sample most of the more formal defined outcomes of the elementary school. Their norms, established so carefully from such large populations, furnish bases of reference not only for the test as a whole but for each of the areas measured. Thus guidance is suggested not only from the results of the test as a whole but also from the results of the single division. And when diagnosis is added to analysis, guidance is greatly facilitated.

These composite examinations help in guiding the transfer student, especially if local norms are available for comparison. As a whole, test batteries (all possessing high reliabilities) are indispensable for testing achievement and for furnishing primary and supplementary data for guidance purposes.

The *major weaknesses of test batteries* lie in the nature of objective examinations themselves. No objective test is able to test the capacity of an individual to gather facts and to marshal them around a problem. They do not test the capacity of an individual to write a theme or essay. In the informational fields of literature, social science, and elementary science there is a strong tendency to ask easily tested questions of information rather than more complicated problems which

TABLE 4. CALIFORNIA ACHIEVEMENT TESTS, ELEMENTARY BATTERY, FORM A

DIAGNOSTIC ANALYSIS OF LEARNING DIFFICULTIES

If the diagnostic profile of a test indicates that a pupil is making normal progress in all fields the teacher will have no use for the following diagnostic analysis. However, where the diagnostic profile shows achievement below a desirable standard in one or more major fields, the following device, which appears somewhere on every Progressive Achievement Test, will assist in identifying and analyzing the specific causes of difficulty as a basis for remedial instruction.

The numerals and capital letters in the diagnostic analysis correspond to the sections of the test similarly marked. For example, if the diagnostic profile shows unsatisfactory achievement in Test 4, Sec. D (addition in arithmetic fundamentals), an inspection of the unsatisfactory responses in this section of the test (by number) will reveal whether or not remedial instruction is needed in carrying, use of zeros, reducing to common denominators, and the like. These topics are then checked by the teacher as the basis for remedial work.

Once an adequate diagnosis has been made, remedial instruction is frequently a simple matter. However, teachers have in the past found the clerical work incident to following each individual pupil a heavy burden. Such extra work is almost completely eliminated if this diagnostic analysis is torn from the test booklet and kept on the teacher's desk where the various items may be checked off as the pupil masters them.

READING

1. Reading Vocabulary

A. WORD FORM:

Lower case words 1-15
 Capitals 16-19
 Miscellaneous type faces 20-25
 (Errors may indicate poor vision)

B. WORD RECOGNITION:

Gross differences 1, 4, 6
 Initial sounds or endings All others
 (Errors may indicate poor hearing)

C. OPPOSITES:

Basic vocabulary 1-23

D. SIMILARITIES:

Basic vocabulary 1-22

2. Reading Comprehension

E. FOLLOWING SPECIFIC DIRECTIONS:

Simple directions 1, 5
 Directions requiring simple choice 2-4, 6-8
 Reading definitions and following directions 9-10

F. INTERPRETATION OF MEANINGS:

Selecting topic or central idea 1, 7

G. REFERENCE SKILLS:

Parts of book 1-2
 Alphabetizing 3-4
 Use of table of contents 5-7
 Use of index 8-10

Understanding directly stated facts 2, 3, 8, 9, 12, 13
 Making inferences 4, 5, 6, 10, 11
 Comprehension of organization of topics 14, 15, 16
 Sequence of events 17-20

ARITHMETIC

3. Arithmetic Reasoning

A. NUMBER CONCEPT:

| | |
|---|-------|
| Writing numbers..... | 1-5 |
| Writing money..... | 6-7 |
| Roman numbers..... | 8-10 |
| Concept of whole numbers..... | 11 |
| Concept of fractions, decimals, and per cent..... | 12-15 |

B. SIGNS AND SYMBOLS:

| | |
|--------------------|------------|
| Signs..... | 1-9, 12-15 |
| Abbreviations..... | 10-11 |

C. PROBLEMS:

| | |
|---------------------------------------|------------|
| One-step..... | 1-4 |
| Two-step..... | 5-8 |
| Sharing and averaging..... | 4, 6, 7, 8 |
| Square measure and cubic content..... | 9, 10 |
| Percentage..... | 13, 14 |
| Ratio..... | 15 |

4. Arithmetic Fundamentals

D. ADDITION:

| | |
|--------------------------|------|
| Simple combinations..... | 1-3 |
| Bridging..... | 4 |
| Carrying..... | 5, 6 |

5. Language

A. CAPITALIZATION:

| | |
|------------------------------|---------|
| First word of sentence..... | 1-3 |
| Names of persons..... | 4, 6, 9 |
| Names of places..... | 3, 6, 7 |
| Days of week and months..... | 5, 9 |
| Abbreviation for months..... | 5 |
| First word of quotation..... | 8 |
| Over-capitalization..... | |

B. PUNCTUATION:

| | |
|-----------------------|--|
| Periods..... | |
| Commas..... | |
| Quotation marks..... | |
| Question marks..... | |
| Over-punctuation..... | |

C. WORDS AND SENTENCES:

| | |
|-----------------|------------|
| Good usage..... | 1, 2, 3, 5 |
| Tense..... | 6, 7 |
| Case..... | 4, 8, 9 |

LANGUAGE

| | |
|----------------------------|-------|
| Number..... | 10 |
| Recognizing sentences..... | 11-20 |

D. SPELLING:

| | |
|-------|--|
| | |
| | |
| | |
| | |

E. HANDWRITING:

| | |
|-----------------|--|
| Legibility..... | |
|-----------------|--|

| | |
|----------------------------------|--------|
| Writing decimals in columns..... | 18, 19 |
| Denominate numbers..... | 20 |

F. MULTIPLICATION:

| | |
|---|--------|
| Tables..... | 1-9 |
| Zeros in multiplicand..... | 2, 5 |
| Zeros in multiplier..... | 7, 8 |
| Two place multipliers..... | 6-9 |
| Mult. with fractions..... | 10, 11 |
| Cancellation of fractions..... | 12, 13 |
| Fractions and mixed or whole numbers..... | 14-17 |
| Pointing off decimals..... | 18, 19 |
| Denominate numbers..... | 20 |

G. DIVISION:

| | |
|--------------------------------------|-----------|
| Tables..... | 1-8 |
| Zeros in quotient..... | 3, 6, 8-9 |
| Remainders..... | 10 |
| Inverting divisors in fractions..... | 11-16 |
| Mixed numbers..... | 16, 17 |
| Reducing fractions to decimals..... | 18 |
| Pointing off decimals..... | 19, 20 |

involve some of the higher thought processes. The application of the ability to use the scientific method or the attitudes of pupils are important questions which remain generally untouched.

This simply means that the test batteries in spite of their length do not tell the whole story. Supplementary facts must be gathered if the whole progress of the child is to be evaluated for purposes of instruction and guidance.

USES OF TEST BATTERIES

The results of testing a school population with achievement batteries may be used in a variety of ways.

The Administrator

The administrator, who must always exercise a broad overview of the total educational situation, finds direct aid in the exercise of his duties from the results of achievement test batteries.

First and foremost, by tabulating the scores and computing the medians of school grades the administrator can observe the average achievement of the different grades and schools of his whole school system. Grade-for-grade comparisons with the norms furnished by the test makers can be made. The administrator's whole program of instruction may be changed in emphasis as a result of grade-equivalent scores derived from these tests.

In the second place, comparisons may be made between the various comparable grades in his system. For example, grade 5 in this manufacturing area may be compared with grade 5 in that better residential area.

In the third place, grade-equivalent scores derived from average achievements in such areas as English, language usage, arithmetic, and geography may focus the need for improvement on certain areas of instruction. Let us say that in this system the pupils from grades 4 to 8 showed decided weaknesses in understanding of what they read. Teacher conferences might then be called, expert teachers of reading called in, and a general program designed to improve achievement in reading for understanding initiated. If, after a reasonable time, alternate forms of the same test were administered and progress in reading assayed then it would be considered that tests had been effectively used.

The Teacher

In some particulars the teacher's aids from tests resemble those of the administrator. He is also interested in the median performance of his class as a whole and in the scores on the various parts. His interest is more specific and more immediate.

In the first place, scores in the different areas of the test acquaint

the teacher with the standing of the class as a whole on each subtest. This is of the first importance. Profile variations of his class as a whole point to both strength and weakness. Especially important is the fact that the general trend is highly dependable. Variations in the same individual might conceivably be attributed to chance errors, but not the general trend. If his class is definitely backward on scores of language usage, this fact can be trusted.

In the second place, profiles of each pupil should be made and studied. Such a graph emphasizes strong and weak points and aids in the acquisition of sound information about the pupil as an individual. Here is, for example, a pupil whose arithmetic scores are good but who is especially low on language and literature. The area needing unusual attention is thus made apparent.

In the third place, teachers can frequently discover more exactly the difficulties in a single area by an item analysis of the test. In a survey of one set of schools with the Metropolitan Achievement Test, we discovered that helpful analyses could be made of the errors which pupils had made on arithmetic fundamentals by means of the following device.

METROPOLITAN ACHIEVEMENT TEST, INTERMEDIATE BATTERY, FORM T
(Analysis of errors)

- I. Items in addition
 - A. Whole numbers: Items 1, 3, 4, 5
 - B. Decimals: Item 40
 - C. Fractions: Items 23, 24, 25, 26
 - D. Zero combinations: Item 2
 - E. Mixed units: Item 52
- II. Items in multiplication
 - A. Whole numbers: 11, 12, 13, 15
 - B. Decimals: 42, 43
 - C. Fractions: 32, 33, 34, 35
 - D. Zero combinations: 14
 - E. Percentage: 54, 55, 56
- III. Items in subtraction
 - A. Whole numbers: 6, 7, 8, 9, 10
 - B. Decimals: 41
 - C. Fractions: 27, 28, 29, 30, 31
- IV. Items in division
 - A. Whole numbers
 - 1. Short division: 16, 17, 18, 19
 - 2. Long division: 20, 21
 - B. Decimals: 46, 47, 48
 - C. Fractions: 22, 36, 37, 38, 39
- V. Graphic presentation: 44, 45, 53
- VI. Changing units of measure: 49, 50, 51, 52

With such details of weaknesses available, substantial changes in materials of instruction and procedures of teaching were made.

The Pupil

In the first place, the objectively scored test may change a pupil's attitude toward his work. The pupil through the instrumentality of the test discovers how he stands in the several areas represented. In so many cases he thinks his standing in a school subject dependent upon the subjective judgment of the teacher. He therefore blames the teacher for his low mark. But this test was not constructed by the teacher, nor do the teacher's whims affect the scoring. With the scoring key in his hand the pupil can check his own paper. The objectivity of the test is a stimulating influence. Under proper guidance he and the teacher go into a huddle and come out with a cooperative plan for the pupil's improvement.

In the second place, the pupil after such an experience may anxiously await a second test which can show him the results of his study. He thus is a competitor with himself, with his own past record.

The low pupil may gain stimulation through considering the norm as a bogey which he may strive to reach. Competition is then not directed toward his peers but toward an impersonal mark set by *other* children—not by the teacher.

Through analyzing the results of his test the pupil may learn to practice more on his own weak points. This attitude may cause the correct result of learning to be achieved with more intensity.

From such considerations it becomes apparent that even test batteries offer many opportunities for improving the ongoing process of education. *It is a pedagogical sin to file the test blanks in such a way that they only gather dust.*

After the batteries of tests have been given and the profiles constructed it is common to find weaknesses in some of the areas tested. There is thus created a need for a more comprehensive test of one area. Among the most important of these areas is that of reading. Reading tests are described in the next chapter.

SUMMARY

The testing program requires the cooperation of all teachers if it is to achieve maximum efficiency. One of the best ways to achieve this cooperation is to enlist their assistance in determining the needs of the school and the particular areas needed to be studied. When the needs have been decided upon and the purposes defined, the selection of the best tests to meet those needs and purposes is undertaken. After the tests are selected, their details of administering, scoring, and inter-

preting must be reviewed with the teachers *before* these activities are undertaken. Most difficult of all for teachers to learn is the process of making records for purposes of interpretation. Following this quantitative and graphical arrangement of records comes the planning of materials and methods for improving conditions found. This is the capstone of the testing program.

The general sequence of achievement tests in a comprehensive testing program is usually (1) the achievement-test battery, (2) the individual subject test, and (3) the diagnostic test. In this text, therefore, achievement-test batteries introduce our discussion of standardized tests. Achievement-test batteries at the elementary level sample rather well the major outcomes of the more formal aspects of education. Since they are standardized on the same population, comparisons may be made between standings in the several subjects of instruction. It makes possible the study of levels of achievement of pupils, classes, schools, and school systems. The achievement levels of pupils may be used to group them within a class and may be highly suggestive of the types of materials suitable for each child's educational progress. For these reasons achievement-test batteries have become customary in American schools.

QUESTIONS AND EXERCISES

1. Plan in considerable detail a testing program for your school. Parallel the description in the text. What aspects of the program do not seem to be included in the text?

2. Discuss the importance of derived scores for purposes of interpretation. Illustrate.

3. Describe in detail the important procedures necessary for administering a test. What does the author mean by a "deadpan"?

4. Explain what is meant by (a) standard score, (b) grade equivalent, (c) percentile score.

5. Describe three graphical procedures usable for interpreting scores.

6. Compare the Stanford Achievement Battery with the Metropolitan

Battery in respect to (a) area covered, (b) establishment of norms, (c) profiles of students, and (d) reliability. Secure samples and manuals and examine them point by point.

7. What are the advantages for education of such tests as the Iowa Every-pupil Tests of Basic Skills and the California Achievement Tests? The disadvantages? To what use can such tests be put in addition to grade placement and subject achievement?

8. To what uses can the administrator put the results of testing? Illustrate.

9. How can the teacher use the results of tests? The pupils?

10. How have the uses of test records for purposes of educational guidance been illustrated in this chapter?

BIBLIOGRAPHY

Books and Manuals

CRONBACH, LEE J.: *Essentials of Psychological Testing*, Chap. 12. New York: Harper & Brothers, 1949.

GREENE, HARRY A., ALBERT N. JORGENSEN, and J. RAYMOND GERBERICH: *Measurement and Evaluation in the Elementary School*, Chap. XXI.

New York: Longmans, Green & Co., Inc., 1942.

HILDRETH, GERTRUDE H., with the collaboration of Harold H. Bixler and the Division of Research and Test Service, World Book Company: *Metropolitan Achievement Tests Manual for Interpreting*. Yonkers, N.Y.: World Book Company, 1948.

Iowa Every-pupil Tests of Basic Skills: *Manual of Interpretation*. Boston: Houghton Mifflin Company, 1940.

KELLY, T. L., GILES M. RUCH, and L. M. TERMAN: *Stanford Achievement Test* (manual). Yonkers, N.Y.: World Book Company, 1940.

Manual, California Achievement Tests for Elementary, Intermediate, and Advanced Tests. Los Angeles, Calif.: California Test Bureau.

Manual of Directions and Interpretations, Gray-Votaw-Rogers General Achievement Tests. Austin, Tex.: The Steck Company.

Master Manual, Coordinated Scales of Attainment, Batteries 1-8. Minneapolis, Minn.: Educational Test Bureau.

PULLIAS, EARL V.: "Commercial Standardized Tests," pp. 65-80, in *Variability in Results from New-type Achievement Tests*, Duke University Studies in Education No. 2. Durham N.C.: Duke University Press, 1937.

TIEGS, ERNEST W., and WILLIS W. CLARK: *Manual of Directions, Progressive Achievement Tests—Advanced Battery*. Los Angeles, Calif.: California Test Bureau, 1943.

TRAXLER, ARTHUR E.: "A Study of

the Revised Edition of the Stanford Achievement Test," pp. 51-57, in 1941 *Fall Testing Program in Independent Schools and Supplementary Studies*, Educational Records Bulletin No. 35, Vol. XIV. New York: Educational Records Bureau, 1942.

———: *Techniques of Guidance*, pp. 75-78. New York: Harper & Brothers, 1945.

WEBB, L. W., and ANNA MARK SHOTWELL: *Testing in the Elementary School*, Chap. XIX. New York: Rinehart & Company, Inc., 1939.

Articles

FORAN, T. G., and M. EDMUND LOYES: "The Relative Difficulty of Three Achievement Examinations," *Journal of Educational Psychology* (1935) 26:218-222.

SPACHE, GEORGE: "Deriving Comprehension, Rate, and Accuracy of Reading Norms for a Short Form of the Metropolitan Achievement Reading Test," *Journal of Educational Psychology* (1941) 32:359-364.

TRAXLER, ARTHUR E.: "Comparison of Scores on the Revised Edition and the Older Edition of the Stanford Achievement Test," *Elementary School Journal* (1942) 42:616-620.

WOOLF, HENRIETTE, and CHRISTINE LIND: "A Study of Some Practical Considerations Involved in the Use of Two Educational Test Batteries," *Journal Educational Psychology* (1935) 26:629-634.

CHAPTER 5

Measurement of Reading, Spelling, and Handwriting

There is some logic in grouping reading, spelling, and handwriting together in one chapter. On many occasions in the elementary and high school they appear in close interrelation, as when a child summarizes in writing what he has read. These three, together with language, constitute the essential tools for further language instruction and for communication. The tests of language are treated in Chap. 6.

READING

In this chapter the section on reading includes a treatment of both elementary and high school tests. The spelling tests described, however, are only those suitable for the elementary school. Spelling tests for the high school are discussed in Chap. 6 under the caption "Language and Literature."

Some authors have considered reading as one of the *receptive* language arts, but reading is certainly more than merely becoming aware of what is on the printed page. Good reading always involves a variety of responses which are related to meaning. Almost all reading-test makers recognize this by affording opportunities to respond correctly to what has been read.

IMPORTANCE OF READING

Learning to read constitutes the major activity of the elementary school. Failure to acquire adequate facility in this process is accompanied with the direst consequences in the upper grades, in high school, and in life. Reading progress needs to be checked at every level of achievement to make certain that satisfactory results have been attained. One of the more difficult problems is to decide upon the optimum time to begin instruction in reading.

OBJECTIVES IN TEACHING READING

As early as 1927 Gist and King stated clearly in a brief statement the major objectives of reading.¹ These frequently quoted aims are:

¹ Gist, A. S., and W. A. King, *The Teaching and Supervision of Reading*, p. 11. New York: Charles Scribner's Sons, 1927. By permission.

- (1) Rich and varied experience through reading.
- (2) Strong motives for, and permanent interest in, reading.
- (3) Desirable attitudes and economical and effective habits and skills.
 - (a) Development of well-established fundamental reading habits.
 - (b) Effective habits of intelligent interpretation.
 - (c) Ability to use books, libraries, and other sources of information economically and effectively.

No satisfactory objective tests have been constructed for Items 1 and 2. If we divide Item 3 into two parts—(a) attitudes, and (b) skills—only the second part is adequately tested. These objectives illustrate the difficulty of constructing tests for goals which are not clearly defined or else defined vaguely.

From the standpoint of defined objectives, the list of reading abilities described by Horn and McBroom (see page 17) gives more promise of successful measurement. They list the abilities (1) to recognize new words, (2) to locate material quickly, (3) to comprehend quickly what is read, (4) to select and evaluate material needed, (5) to organize what is read, and (6) to remember what is read. Their listing furnishes us with definite, measurable objectives. Their last three abilities, including attitudes toward the care of books and toward attacking reading with vigor, together with the knowledge of sources, offer few opportunities for developing satisfactory measuring instruments.

TESTS OF READING IN ACHIEVEMENT-TEST BATTERIES

Reading tests constitute an integral part of all achievement-test batteries. Generally speaking, there is a set of paragraphs of increasing difficulty whose comprehension is tested by the completion technique, as in the Metropolitan and Stanford achievement tests, or by multiple-choice items, as in the California and Coordinated tests. In most tests there are also vocabulary tests which may be answered either by recognizing what a word means or by giving its opposite. In those batteries which specialize on the measurement of reading, arithmetic, and language many more details are possible, and a test quite comparable to tests devoted entirely to reading is achieved. Illustrations of such tests are (1) the California Achievement Tests, and (2) the Iowa Every-pupil Tests of Basic Skills.

The California Achievement Tests go so far as to issue a separate manual for reading, which indeed may be treated as a test separate from the battery as a whole.

The reading test for grades 7, 8, and 9, for example, includes tests

of vocabulary, information about a book, the use of an index, and tests of understanding paragraphs. The 90-word vocabulary test, answered by giving opposites to words, is divided into four parts equally distributed among words needed in (1) mathematics, (2) science, (3) social science, and (4) general reading. Their tests on reading comprehension deal with the ability to follow directions and, as the manual puts it, "The test situations measure the students' ability to (1) read and comprehend directly stated facts, (2) select the best topics or central ideas, (3) make inferences and deductions from written material, and (4) read and comprehend the author's ideas as expressed in paragraphs" (page 3). Reliabilities of the two parts and of the reading test as a whole are about .90. It also furnishes details for constructing a diagnostic profile.

The reading test of the Iowa Every-pupil Tests of Basic Skills is called Test A: Silent Reading Comprehension. The advanced battery for grades 5, 6, 7, and 8 is divided into Part I, Reading Comprehension, and Part II, Vocabulary.

The 50-word vocabulary tests includes many words suitable for these grades. Such words as "desirable," "indefinite," "civil," and "essential" are set in multiple-choice items. The tests of comprehension, which are of the work-study type, are excellent examples of test construction. In the first place the selections are much longer than usual, each one filling a large page and consisting of three to five paragraphs. In addition, their subject matter is concerned with material little known to the reader. Such selections as "The Boomerang," "The Seiche," "Billy Sunday," and "The Northwest Passage" constitute new material for most readers in the upper grades. Some of the questions are informational, with the answers contained in the paragraphs, but many of them call for understanding and interpretation. One example must suffice: in one paragraph there is a verbal description of the shape of a boomerang but the question asks the reader to select from four visual shapes the one "most nearly the shape of a boomerang." The author, in his search for tests of reading comprehension, has been unable to find another test as good as this one. It tests (1) the meaning of words, (2) the meaning of sentences, (3) the meaning of paragraphs, and (4) the relation among paragraphs.

The present chapter considers (1) tests of reading readiness, (2) tests of reading achievement, and (3) tests of reading diagnoses.

Tests of Reading Readiness

While intelligence tests are of value for guidance in beginning formal instruction in reading and number work, they do not predict final achievement marks as well as tests of reading readiness. There was

needed an instrument which would test specifically those traits on which instruction in reading depends. It is these instruments which are now to be described.

Reading-readiness tests were constructed to measure precisely those traits which are required to learn to read. Careful analyses were made of those traits which reflected clearly the maturing process. Among these traits were the following:

1. Language growth
2. Correctness of language usage
3. Interest in learning to read
4. Visual and auditory discrimination and reasoning ability
5. Knowledge of facts and events in common experience
6. Number information
7. Motor control
8. Ability to pay attention to and understand simple stories.

Most of the good tests of reading readiness have been based on recognition of and understanding of these processes. The following are usually recognized as good tests of reading readiness:

1. Metropolitan Readiness Tests
2. Gates Reading Readiness Tests
3. Lee-Clark Reading Readiness Test
4. Reading Aptitude Tests by Marion Monroe
5. Stevens Reading Readiness Test

One of these will be described at some length as a sample, and there will follow a discussion of the value and use of reading-readiness tests in teaching and guidance.

The Metropolitan Readiness Tests attempt to sample the majority of the traits known to be characteristic of readiness to read.¹ There are six tests in the battery, with additional information gained from the drawing of a man and from writing one's name.



























Test 1 measures visual discrimination by means of a series of paired figures, some of which are alike and some different. The material varies from two boats (like), to an ellipse and a circle, to pairs of one-place, two-place, and three-place numbers (Fig. 5). Moreover, the child may recognize likenesses and differences between pairs of two-letter, three-letter, four-letter, and five-letter words.

Test 2 involves the copying of 11 figures of varying degrees of complexity. Such figures as a circle, a square, a diamond, and a swastika are used. Then in the test come materials to be copied which are much like actual schoolwork. These are the letter N, an h, 63, C.A., and SDL.

Test 3 is a test of vocabulary. Nineteen words given orally are to be

¹ Items from these tests quoted by permission of World Book Company, Yonkers, N.Y.

TEST 1. SIMILARITIES

| | | | |
|---|---|---|---|
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  | 2 | 2 |
|  |  | DC | CD |
|  |  | 35 | 35 |
|  |  | GA | GA |
|  |  | 291 | 216 |
|  |  | on | no |
|  |  | boy | boy |
| | | flag | flies |
| | | chick | chair |
| | | threw | threw |

Rights minus wrongs . . . (Score, Test 1)

FIG. 5. Metropolitan Reading Readiness Test.

recognized from drawings, each word from four drawings. Words such as "key," "desk," "bridge," "jewel," "blossom," "bonfire," "insect," and "poultry" are used. Words like "moccasin," "chariot," "insect," and "poultry" are at the more difficult end of the scale. Fig. 6 shows sample pictures. In No. 8 the word is "lantern"; in No. 12, the word is "melon"; and in No. 17, the word is "insect."



FIG. 6. Metropolitan Readiness Tests, Vocabulary, Items 8, 12, and 17.

Test 4 carries on the idea of understanding of words but now the words are combined into sentences. From among four pictures the child must recognize "The pumpkin in the window," or "The man is reading a book," or "The man at the drugstore has things we need. He sells medicine and things for sick people."

Test 5 is a very complete sampling of the number information possessed by children. Four pages of pictures and 40 questions about numbers and their meaning are given. The child is asked about length and breadth ("Mark the widest board"), the recognition of a circle and triangle from their names, to write numbers such as 5, 9, and 6, to understand how to count seven and thirteen, to know something of

TABLE 5. TESTS OF READING READINESS

| Test | Reliability | Achievement correlations | Intelligence-test correlations | Contents |
|--|--------------------------|---|--|--|
| Gates Reading Readiness (group), Teachers College, Columbia University | .97 (<i>N</i> , 174) | Gates Primary Achievement test = .706 | Pintner-Cunningham (Gates Reading Readiness + Pintner-Cunningham) = .76 | Test 1. Picture directions Test 2. Word matching Test 3. Word-card matching Test 4. Rhyming Test 5. Reading letters and numbers |
| Metropolitan Readiness Tests (group), World Book Company | .83 to .89 | | *Pintner-Cunningham = .53 (<i>N</i> , 94); Detroit First Grade Intelligence Test = .70 (<i>N</i> , 34) *Combination of 3 intelligence tests, .79 | 1. Recognition of likeness and difference between forms and letters 2. Copying figures 3-4. Comprehension of words and phrases 5. Number knowledge 6. Common knowledge |
| Stevens Reading Readiness (group), World Book Company | .96 | Teachers' Ratings of Achievement = .80 (<i>N</i> , 460) after 70 days' reading instruction | | 1. Recognition of objects and letters different from among others 2. Recognition of words and phrases from among others |
| Lee-Clark-Reading Readiness Test (group), California Test Bureau | .92 (<i>N</i> , 170) | Lee-Clark Reading Tests (primer) = .67 | California Test of Mental Maturity = .65 | Test 1. Matching letter symbols Test 2. Crossing out letters different from others Test 3. Vocabulary and following instructions Test 4. Identification of letters and words |

TABLE 5. TESTS OF READING READINESS (*Continued*)

| Test | Reliability | Achievement correlations | Intelligence test correlations | Contents |
|---|-------------|--|--------------------------------|--|
| Monroe Reading Aptitude Tests (partly group, partly individual), Houghton Mifflin Company | .87 | Gray's Oral Reading Test and Iota Word Test = .75 (<i>N</i> , 85) | | Group Tests 1. Visual <i>a.</i> Identifying forms and their positions <i>b.</i> Tracing a maze <i>c.</i> Drawing a picture 2. Motor <i>a.</i> Dots in circles <i>b.</i> Keeping on a line with pencil 3. Auditory <i>a.</i> Recognizing correct pronunciation <i>b.</i> Recognizing a word sounded out phonetically 4. Vocabulary |
| Betts Ready to Read Battery of Tests (individual), Psychological Corporation | | | | Preschool through college. Physiological and psychological tests |
| Van Wagenen Reading Readiness Tests (individual), Educational Test Bureau | .94 | Reading tests, end of grade 1 = .73 | | 1. Range of information 2. Perception of relations 3. Vocabulary 4. Word discrimination 5. Memory span for ideas 6. Word learning |

simple ordinal numbers, to recognize $\frac{1}{2}$, and to do the simplest subtraction and addition.

Test 6, a test of information, asks questions which involve the recognition of common objects among four pictures. The child is asked to mark "the thing to carry when it rains," "what helps people to see better," and "the thing in which to go across the ocean."

Test 7 is the problem of drawing a man and of writing one's name.

As one reads the content of these tests it is evident that they contain problems more precisely like reading than does the general intelligence test. They have the great advantage of being analytical and of furnishing details of weaknesses in certain areas. Weakness lies, let us say, in vocabulary or in visual discrimination or in the combination of words. Teaching then can be directed toward the points of weakness and at the level of achievement. Percentiles are furnished for each test. From this test, inferences as to when to begin reading can be drawn from the test as a whole as well as to the area and magnitude of the weakness which prevents the child from being ready for formal reading.

The chances for success in reading are calculated for each level of scores received on the Metropolitan Readiness Test, and a critical score is furnished below which the chances of success are small indeed. Since many other factors enter into success besides what can now be measured, this notion of chances of success aids the teacher in making tentative any grouping of children based on the scores received in the test.

Intelligence tests and tests of reading readiness together furnish information highly predictive of subsequent success in reading. The latter tests especially break down the total aptitude for reading into special areas where modifications in programs of materials and procedure can be made. Definite conclusions can thus be drawn as to whether or not a child is ready to begin formal instruction in reading. Guidance of the finest kind can thus be rendered at the very beginning of a school career.

Table 5 contains a list of tests of reading readiness.

Tests of Reading Achievement

Achievement tests in reading offer a much larger sampling and a more complete coverage of the wide variety of reading situations than is possible in the test battery. In the usual testing program the test battery is given first. From the test records thus obtained unsatisfactory results might be discovered in any one of the tested areas: reading, arithmetic, language, etc. Let us suppose that one of these unsatisfactory areas is reading. Before undertaking a program intended for improvement of the children's reading abilities it is best to give a more comprehensive reading test. From such a test, there may be obtained (1) a more dependable report of the children's general reading abilities, and (2) more analysis of the difficulties which the children have in reading. In the secondary school where test batteries have not been too satisfactory, achievement tests of reading have been of very great value. It has been discovered that poor scholarship in several

subjects has frequently been due to the students' failure to form satisfactory reading habits in the elementary school.

Reading Tests in the Elementary School

The importance of reading has been so universally acclaimed that a large number of instruments have been constructed for its measurement. Gray's tests of oral reading, described on page 108, have proved

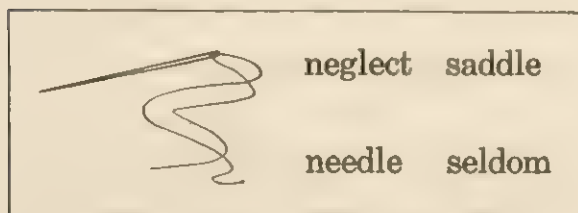
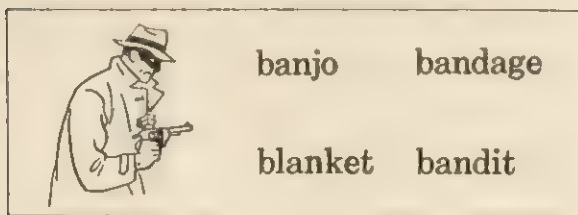
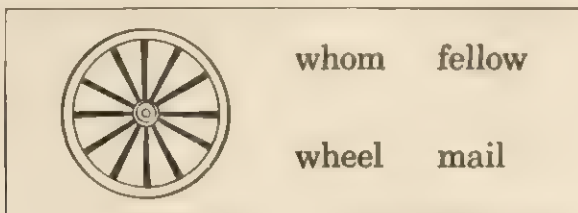


FIG. 7. Gates Primary Reading Tests, three items.


so satisfactory for this purpose that few competitors have appeared. In the realm of silent reading, on the other hand, literally dozens of tests have been published. A selected list appears on page 113.

Primary—Grades 1, 2, and 3. Instruction in the primary grades consists largely of adapting suitable materials to the levels of growth discovered by the teachers' observations and by scores from intelligence tests and from tests of reading readiness. Achievement tests of value during grades 1 and 2 are largely reading tests.


A good illustration of a test of achievement which also offers various opportunities for diagnosis and guidance is the Gates Primary Reading

Tests.¹ These tests are divided into three types: (1) word recognition, (2) sentence reading, and (3) paragraph reading.

In Type 1, the test of word recognition, a picture of an object is presented in the left part of a drawn box. In the right section there are four printed words, one of which is the name of the picture. The directions say, "I want you to look at the first picture. Next to it there



7. Do you like to go camping? It is fun to sleep in a tent. Draw a line under something you might take on a camping trip.



13. A pumpkin with a funny face stands for Hallowe'en and a lighted tree for Christmas. Easter brings Bunny with her basket of eggs. Put X on what stands for Hallowe'en.

FIG. 8. Gates Primary Reading Tests, short sentences, Items 7 and 13.

are some words. One of the words goes with the picture. You are to draw a ring around that word that tells about the picture." Figure 7 shows an illustration from the test.

There are altogether 48 items, and 15 minutes is allowed for the test. Such words as "sit," "hen," "bear," "clock," "stand," "crow," "pick," "window," "leaf," "lake," "roof," and "drive" are included.

The second part of this test, Type 2, contains short sentences printed

¹ Items used by permission of Bureau of Publications, Teachers College, Columbia University, New York.

out with appropriate answers indicated in pictures. Figure 8 gives two illustrations.

The sentences increase in length and complexity from "This is a hat," to "This bottle is full of ink," to "The young daughter has pretty clothes." There are 35 items, and 15 minutes is allowed for taking the test.

The third part of the test, Type 3, paragraph reading, is built along the same lines as the first two but has longer passages to read and

SET I—No. 2

A boy had a dog.
The dog ran away.
The boy ran after him.
He ran very fast.
He caught the dog.
He took him home.
The boy said,
"You are not a good dog.
You must stay at home."

interpret. The complexity of the paragraphs vary from "Draw a line under the long train," to "A mother told her boy to jump into the car and stay there. Draw a line from the boy to the car," to a paragraph made up of three complex sentences. Twenty minutes is assigned for taking this section of the test.

From these descriptions and illustrations it is clear that many opportunities for analysis of reading habits arise. A child might be weak in vocabulary, or he might be acquainted with words but be unable to interpret them in a sentence, or finally he might understand sentences written singly but be unable to interconnect several ideas in a paragraph.

This test of primary reading has satisfactory reliability ($r = .92$) and has been checked against teachers' estimates of word recognition, sentence meaning, and paragraph understanding.

SET II—No. 1

A nest is in a big green tree. The mother bird made the nest. She put it on the branch of the tree among the pretty leaves. She made it of twigs, leaves, and grass. She put soft rags inside of it. The nest has five baby birds in it.

The nest is large and round. The little birds will not fall out. The nest holds the mother bird and the little birds, too. It is hidden under the leaves. The old cat cannot see it. He does not know where the birds are. He will not find them there.

The nest is the home of the birds. It is a bed for the baby birds. The wind rocks it back and forth. The nest is very strong and the wind cannot blow it down. The little birds eat and sleep all day. They will learn to fly very soon.

Another test useful for checking the child's perception of words and phrases is the Detroit Word Recognition Test. Forty pictures are presented, each of which is named or described by a word or phrase. The problem is to draw a line from the word or phrase to the proper picture. The words were selected with great care from Thorndike's *Word Book* so that usefulness for work in the elementary school was assured. Norms are furnished for tests given both at the beginning

and at the end of the terms, as well as for the bright, medium, and dull. The reliability of the test seems to decrease with the grade. From grade 1B to 2A these figures are .86, .77, .72, and .52. It is clear that it would be of most value in grade 1. The correlation with teachers' estimates of ability to recognize words is .74.

The third instrument useful for success in reading in the first two grades is Gray's Oral Reading Test. This test as a whole consists of a set of selections suitable for oral reading in grades 1 through 8. There are four sets:

Set I. First grade

Set II. Second and third grades

Set III. Fourth and fifth grades

Set IV. Sixth, seventh, and eighth grades.

Each set has five samples of approximately equal difficulty. Those illustrated on pages 106 and 107 are Set I, No. 2, and Set II, No. 1. While the pupil reads aloud, the tester keeps a record. The procedure recommended in Gray's Manual for recording the results of the reading is shown in the accompanying illustration and instructions.

The sun ^{may} pierced into my large windows. It was the opening of October, and the ^{clear} sky was of a dazzling blue. I looked out of my window and down the street. The white houses of the long, straight street were almost painful to the eyes. The clear atmosphere allowed full play to the sun's brightness.

If a word is wholly mispronounced, underline it as in the case of "atmosphere." If a portion of a word is mispronounced, mark appropriately as indicated above; for example, "pierced" pronounced in two syllables, sounding long a in "dazzling," omitting the s in "houses," the al- in "almost," or the r in "straight." Omitted words are marked as in the case of "of" and "and"; substitutions as in the case of "many" for "my"; insertions as in the case of "clear"; and repetitions as in the case of "to the sun's." Two or more words should be repeated to count as a repetition.

Adequate norms are furnished. Most useful of all is the opportunity for the tester to observe the child's pronunciation, his attempts to recognize words, his omission of letters in words, and his tendencies to supply words which do not appear in the paragraph. From such analytical observations, diagnoses of difficulties can be secured and guidance furnished just at the point where it is needed.

Intermediate—Grades 3 to 8. There are a great many tests suitable for testing reading in the intermediate grades. Among these, two are selected for description: The Gates Tests of Silent Reading (grades 3 to 8) and the Iowa Silent Reading Tests, elementary test (grades 4 to 9).

Gates Silent Reading Tests

The Gates Silent Reading Tests are divided into four parts:¹

Type A, Reading to Appreciate General Significance consists of a set of 24 short paragraphs, to be read for 6 minutes for accurate general impression. The reading required resembles that used in casually reading a novel or newspaper. An item of about medium difficulty is used here for illustration:

10. The dog ran to meet the man coming up the path. He wagged his tail joyously and barked with short, excited barks. The man leaned down and patted the dog on the head. Then he rolled up the paper that was under his arm and gave it to the dog. The dog ran with it up the path toward the house, his tail wagging all the time.

Draw a line under the word that best tells how the dog felt: sad afraid
lonely weary happy

Type B, Reading to Predict the Outcome of Given Events, is also composed of 24 items, to be read for 8 minutes. This kind of reading involves analysis of what is read and a thinking of the facts together to predict the outcome of the events described. An example is:

11. Pat Dolan lived in a crowded part of New York City. His parents were very poor. What money he earned selling papers he gave to them. One day a woman gave him a quarter. Pat had always longed to ride on a big green bus. He could hardly wait until Sunday when he did not have to go to school or sell papers. At last Sunday came.

Pat bought a toy dog with a squeak
He went to church in his father's car
He took a long ride on a big bus
He sold a hundred papers that day

Type C, Reading to Understand Precise Directions, consists also of 24 items, to be read for 8 minutes. All items contain pictures which are marked in some way indicated in the paragraph. The correct reading of these paragraphs involves "rigid, careful reading" (Fig. 9).

Type D, Reading to Note Details, includes 18 items, to be read for 8 minutes. The reader must comprehend several points in a paragraph at once. A sample follows:

10. In the mountains we find many pretty flowers. Among those that can be found in the early fall are the goldenrod and purple aster. Think of the color they give to the side of the hills. A story tells that these two flowers were once two little girls who wanted to make everyone happy. So a fairy changed them into goldenrod and asters.

¹ Items used by permission of Bureau of Publications, Teachers College, Columbia University, New York.

When are goldenrod and asters found?

Spring Summer Fall Winter


What does the story say these flowers were once upon a time?

Stars Girls Sunbeams Boys


How did they want to make everyone feel?

Gay Excited Young Happy

Inspection of the characteristics of the four types of reading will convince anyone of the close relationship between objectives in teaching



11. Some things grow on trees and some things grow in the ground. Here is an apple, a walnut, a banana, and a beet. Apples, walnuts and bananas grow on trees, and beets grow in the ground. Draw a line under the ones that grow on a tree.



17. The middle part of this bridge is a draw-bridge over a river. It is raised to let the ships go through and closed to let the trains go across. Make a cross on the part of the bridge that will be raised up when a ship gets near.

FIG. 9. Gates Silent Reading Tests, Type C. Reading to understand precise directions.

reading and these measuring instruments. The tests are easily scored for each of the four types -A, B, C, D. Since norms are furnished both for the test as a whole and for the four parts, an individual's reading score for the test as a whole and for each part can be interpreted. Thus a student may be high or low on all four types or high in some and low in others. In this manner some analysis can be made of the pupil's

difficulties in reading. A pupil also may try only a few items but get them all right, or he may attempt many with only a small number correct, or he may follow some middle course. We may thus discover a slow, laborious reader, a rapid, haphazard sort of a reader, and one who reads at a normal rate with normal success. All these facts aid the teacher in her attempt to provide materials and procedures for improving the reader process. Gates has furnished suggestions and further reading for improving poor reading as indicated by scores in each of the four types of the test.

The Gates Silent Reading Tests have been widely used. There are, however, three important limitations of the tests. In the first place there is no experimental evidence that the four types of tests actually measure the outcomes of instruction. In the second place, the paragraphs are short and in many cases too easy for children of the upper grades. Since there is little gradation of difficulty the test tends to become a rate test. In the third place, the types are certainly not independent. Table V of the manual shows correlation ranging from .66 to .92 between scores received from the different types. The correlations of Type A with Type B are above .80 in 14 out of the 15 coefficients. Such high correlations indicate a tremendous amount of overlapping between the types.

The Iowa Silent Reading Tests, elementary form, are suitable for grades 4 to 9. It made use of the objectives described by experts in the field to build a test which would reflect satisfactorily improvement

TABLE 6. IOWA SILENT READING TESTS: ELEMENTARY TEST
Reliability

| | |
|--|---------------------|
| Test 1. Rate and comprehension..... | .83 (rate) |
| Science material | .68 (comprehension) |
| Social-studies material | |
| Test 2. Directed reading..... | .92 |
| Science material | |
| Social-studies material | |
| Test 3. Word meaning..... | .86 |
| General vocabulary | |
| Subject-matter vocabulary | |
| Test 4. Paragraph comprehension..... | .85 |
| Selection of central idea of paragraph | |
| Identification of details essential to the meaning of the paragraph | |
| Test 5. Sentence meaning..... | .60 |
| Test 6. Location of information | |
| Alphabetizing; using guide words..... | .94 |
| Use of index..... | .81 |
| Median standard score..... | .93 |

in each objective (see page 17, Chap. 2). An element of strength is its four equivalent forms. Tests for most of the objectives as described by Horn and McBroom are provided. The divisions of the test, with reliabilities based on 220 cases in grade 6, are as shown in Table 6. This test of silent reading of the work-study type includes passages of four to five paragraphs in length to be read in the section on directed reading. In the section on paragraph comprehension two sorts of questions are asked: (1) on selecting the topic of the paragraph, and (2) on

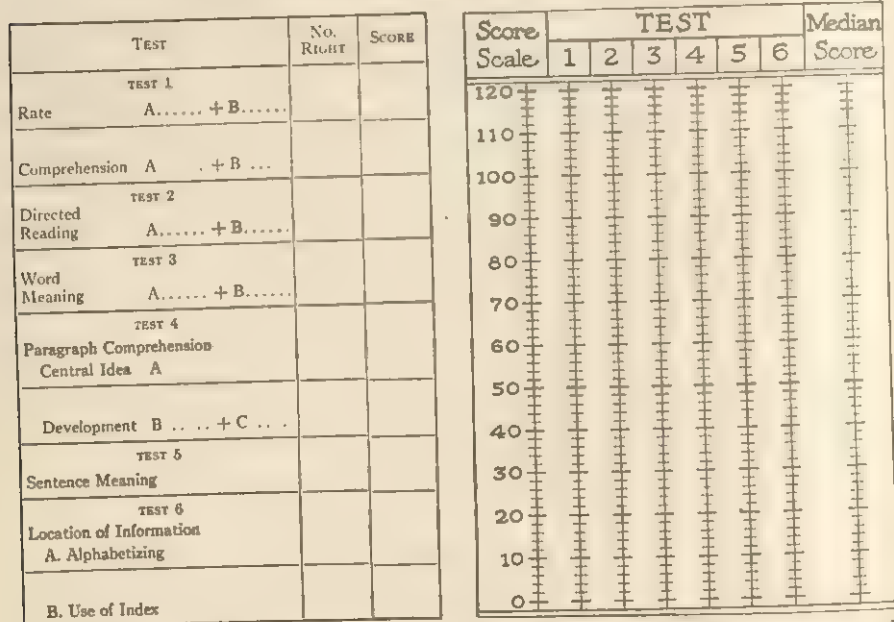


FIG. 10. Profile chart, Iowa Silent Reading Tests (elementary). (By permission of World Book Company.)

the contents of the paragraph. All scores are transmitted to standard scores, and a profile chart is then filled in (Fig. 10). In interpreting such a chart, attention must be directed to (1) the reliability of each test and (2) the intercorrelations of the tests. Table 6 indicates that very little dependence for an individual's score can be placed on Test 5, since its reliability is only .60, or on the comprehension score in Test 1 (reliability, .68). In the case of the other tests, the reliabilities range from excellent (.94) to fair (.81). In the second place, the correlations of the tests with each other are satisfactory. They range from .18 to .63 with most of the correlations in the .20's, .30's, and .40's. This low intercorrelation indicates that there is only a small degree of overlapping between the various tests, and therefore each test may be thought of as testing aspects of reading different from the others. Norms have been

carefully computed from 1,600 to 1,900 cases gathered from 19 widely different communities located in 13 states.

In comparing the Gates Silent Reading Tests and the Iowa Silent Reading Tests, both widely used, we find the Gates less difficult. This test is less rigidly standardized than the Iowa and much of its reading is of short paragraphs. It has an indication of rate in the number of items attempted and a fine score of accuracy. The Iowa test, on the other hand includes sentence meaning and the location of information not contained in the Gates. There is less overlapping among its constituent tests. It, however, does not have reading for prediction. There is little difference in the reliabilities of the tests which compose the total. Either test can be profitably used, with a preference for the Gates test in grades 3 and 4, and a vote for the Iowa test in grades 5 to 8. Included here is a selected list of reading tests suitable for the elementary grades.

LIST OF ACHIEVEMENT TESTS IN READING FOR THE ELEMENTARY SCHOOL

1. Iowa Silent Reading Test, elementary, grades 4-9. World Book Company, Yonkers, N.Y.
2. Detroit Reading Tests, grades 2-9. World Book Company, Yonkers, N.Y.
3. Gates Silent Reading Test, grades 1-2. Teachers College, Columbia University.
4. Los Angeles Primary Reading Test, grades 1-3. California Test Bureau, Los Angeles, Calif.
5. Emporia Silent Reading Test, grades 3-8 (survey). Kansas State Teachers College, Emporia, Kans.
6. Traxler Silent Reading Tests, Series I, grades 7-9. Public School Publishing Company, Bloomington, Ill.
7. Monroe Revised Silent Reading Tests, grades 3-8. Public School Publishing Company, Bloomington, Ill.
8. Sangren-Woody Reading Test, grades 4-8 (survey and diagnostic). World Book Company, Yonkers, N.Y.

Separate tests of arithmetic, of social science, and of language appear in the chapters on the measurement of mathematics, of social science, and of English respectively.

Reading Tests at the High School Level

There are many reading tests at the high school level. Among these, three will be mentioned:

1. Iowa Silent Reading Tests (Advanced), grade 10 and above. World Book Company, Yonkers, N.Y.
2. Traxler Silent Reading Tests, Series II, grades 10-12. Public School Publishing Company, Bloomington, Ill.
3. Cooperative Reading Comprehension, grades 7-12. Educational Testing Service, Princeton, N.J.

The Iowa Silent Reading Tests have been widely used. The advanced test is composed in the same manner as is the elementary. The divisions of the tests with their reliabilities appear in the following table.

| | |
|---|---------------------|
| Test 1. Rate and comprehension..... | .73 (rate) |
| Science material | .82 (comprehension) |
| Social-studies material | |
| Test 2. Directed reading..... | .91 |
| Test 3. Poetry comprehension..... | .80 |
| Test 4. Word meaning..... | .90 |
| Social studies | |
| Science | |
| Mathematics | |
| English | |
| Test 5. Sentence meaning..... | .85 |
| Test 6. Paragraph comprehension | |
| Selection of central idea of paragraph..... | .54 |
| Identification of details essential to the meaning of the paragraph..... | .73 |
| Test 7. Locating information | |
| Use of index..... | .82 |
| Selection of key words..... | .91 |

The prose passages to be read are composed of four to eight paragraphs. While there is opportunity for asking questions involving the interrelations of paragraphs, none is asked. An interesting feature is the selection of words from four areas in the vocabulary test. From social-science material, 20 words are to be defined including such words as "capital," "suffrage," "contraband," and "amnesty." Samples from the science vocabulary of 15 words are "density," "adhere," and "latent." The mathematics vocabulary is composed of 15 terms, of which "degree," "origin," and "linear" are samples. The fourth division of vocabulary is made up of 20 English terms such as "legend," "allegory," "satire," and "epigram."

The areas of the test were selected after a careful study of materials and objectives by Horn and McBroom (see page 17). Norms were computed from over 10,000 cases. Its raw scores are transmuted immediately into standard scores, and these, by means of a table, are turned into percentiles. Provision is made on the front page of each test for constructing a reading profile of the seven divisions of the test. This test is a satisfactory silent-reading test of the work-study type. The material is somewhat academic and the techniques a trifle artificial. The test has the possibility of rewarding too highly the rapid reader.

Tests of Reading Diagnoses

Programs of testing usually begin with a test battery, proceed with a more complete coverage of a single area, and end with a diagnostic test. Diagnostic tests are so arranged that weak points are discovered and small errors defined whose correction produces greater accuracy

and understanding in the material read. When these critical areas are discovered programs of study aimed at a narrower function may be prepared. By this procedure more rapid progress is assured.

All analyses of reading difficulties stem from an understanding of the reading process itself. This process is much more complicated than at first appears. On the one hand, it depends upon clear perception of the symbols involved; on the other, it depends on the association and relation of these symbols among themselves as well as their relations to life experience. Good reading involves the formation of a hierarchy of habits which unifies several words into one idea, but the correct idea depends upon the accurate perception of the words themselves. Unless the words are accurately perceived reading may become an imaginative procedure in which words are added or subtracted with impunity. This nicety of balance between analysis and synthesis must ever be maintained.

In the case of many children who become poor readers, failure in the clarity of perception, in word analysis, or in that unity of parts from which meaning derives is apt to take place. In some diagnostic reading tests, the emphasis is on perception; in others, on the relations of words; while others attempt to discover the point where the lack of meaning or recall appears. A few tests attempt to make a complete analysis of the individual's reading, with enough samples at each critical point to determine the exact location of the difficulty.

The Durrell Analysis of Reading Difficulty is a diagnostic test for grades 1 to 6.¹ It presupposes (1) a medical record which contains a careful check of the efficiency of vision and of hearing, and (2) pertinent facts gathered from the home which relate to possible sibling rivalry, change of the dominant hands, emotional reactions, and special interests. In the second place, this analysis of reading furnishes a check list of difficulties which is quite inclusive. The following are samples from the check list, which may be regarded as a diagnostic record sheet:

1. Background skills

Hearing vocabulary poor

Hearing comprehension poor (determined by previous test)

Faulty voice or speech habits

2. Word mastery skills

Word recognition

Low sight vocabulary

Will not try difficult words

Can spell but not pronounce

Ignores word endings

Guesses at words from general form

¹ World Book Company, Yonkers, N.Y. Items by permission.

3. Word analysis

Word analysis ability poor

Will not try difficult words

Has no method of word analysis

Sounds aloud by single letters blends- syllables

Unable to combine sounds into words

Looks away from word after sounding

Sounding slow or inaccurate

Spells words: successful—inadequate

Silent word study: successful inadequate

Enunciates badly when prompted

Systematic errors (tabulation of them)

Names of letters not known

Sounds of letters not known

Blends not known

In like manner check lists are included for analysis of difficulties in oral reading (both phrase reading and comprehension) and in general reading habits. In silent reading the check list is unusually complete. There is included a check list of mechanics as follows:

1. Low rate of silent reading
2. High rate at the expense of mastery
3. Lip movements: constant—occasional
4. Whispering: constant—occasional
5. Lacks persistence in hard material
6. Marked insecurity evident
7. Poor attention necessitates rereading.

Similar check lists are provided for comprehension, eye movements, comparison with oral reading in speed, recall, and security, in oral recall, in written recall, in study skills, in spelling, and in writing.

The test itself is divided into paragraphs for oral reading with questions after each paragraph. Furthermore, there are paragraphs in a cardboard manual to be read orally and then recalled. In the record blank there are phrases to be checked whenever they are recalled and widely spaced lines for recording any errors. During the recall there is no aid to be given. You simply say, "Tell me everything that you can remember of that story." In a like manner paragraphs are read silently and then their content is recalled.

A small tachistoscope has been constructed by means of which one word from a list may be exposed for a short length of time. There are two parts in each of four lists. Part 1 is meant to study flash recognition, while Part 2 is to study the analysis of words. All incorrect responses are recorded phonetically. Finally there is a phonetic inventory, a

place for recording difficulties in spelling, difficulties in handwriting, and difficulties in written recall.

Standards for the various parts of the test based on approximately 1,000 children are furnished. The author regards the opportunity for observation of errors under standard conditions as much more important than the norms. The check list of errors was based on the errors discovered in the reading of 4,000 children brought to the clinic. The manual states, "The check list of errors will be found to include all of the significant errors made by any child."

This test lacks a thoroughgoing study of its reliability. In some instances the norms are not entirely clear. For example, the norms for written recall in silent reading are the same as those for oral recall. The author suggests that these two norms are sufficient for rough analysis. One student (Miles A. Tinker) thinks that the items concerning eye movements are of dubious value. On the whole, though, the critics agree that this instrument provides an especially helpful instrument for diagnosing and recording specific difficulties in reading.

Tests of Oral Reading

Tests of oral reading, necessarily given individually, offer an opportunity both to check the level of achievement in this function and to diagnose reading difficulties present in both silent and oral reading.

One of the few oral reading tests, Gray's Oral Reading Test is definitely a diagnostic test. The tester makes his own records (1) in seconds, for each paragraph, and (2) in notes written on each paragraph (see page 108 for illustration).

The best use of the test appears in the study and classification of errors which are made. Ordinarily No. 1 of a set is given. Then after two or three weeks, No. 2 is given. In the meantime attempts are made to provide the sort of training which will produce improvement. However, records of errors made in informal reading are also kept and the total errors entered on the "individual record sheet" (Fig. 11).

It is clear from the study of this sheet that a satisfactory analysis of the mechanics of oral reading can be made. There is, however, no record of reliability in the manual. In a diagnostic test this is not as important as in other tests. The test's only major weakness is a failure to check the comprehension in any way.

In Table 7 appears a partial list of diagnostic reading tests.

SPELLING

The outcomes of the teaching of spelling have been clearly defined during the last half century. The number of words to be learned have been greatly reduced. Instead of the vast number of words, both usual

INDIVIDUAL RECORD SHEET

Progressive Analysis of Errors in Oral Reading

| Pupil's Name _____ | | Age _____ | | Grade _____ | | | | | | | |
|---|--|-----------|-------|-------------|-------|-------|-------|-------|-------|-------|-------|
| Types of Errors | | No. 1 | Daily | No. 2 | Daily | No. 3 | Daily | No. 4 | Daily | No. 5 | Daily |
| I. INDIVIDUAL WORDS | | | | | | | | | | | |
| 1. Non recognition | | | | | | | | | | | |
| 2. Gross mispronunciation | | | | | | | | | | | |
| 3. Partial mispronunciation | | | | | | | | | | | |
| a. Monosyllabic Words | | | | | | | | | | | |
| 1. Consonant { Initial | | | | | | | | | | | |
| { Middle | | | | | | | | | | | |
| { Ending | | | | | | | | | | | |
| 2. Vowel { Initial | | | | | | | | | | | |
| { Middle | | | | | | | | | | | |
| { Ending | | | | | | | | | | | |
| 3. Consonant blends { Initial | | | | | | | | | | | |
| { Middle | | | | | | | | | | | |
| { Ending | | | | | | | | | | | |
| 4. Vowel digraph { Initial | | | | | | | | | | | |
| { Middle | | | | | | | | | | | |
| { Ending | | | | | | | | | | | |
| 5. Pronounce silent letters | | | | | | | | | | | |
| 6. Insert letters | | | | | | | | | | | |
| 7. Pronounce backwards | | | | | | | | | | | |
| 8. Rearrange letters | | | | | | | | | | | |
| b. Polysyllabic Words | | | | | | | | | | | |
| 1. Accent | | | | | | | | | | | |
| 2. Syllabication | | | | | | | | | | | |
| 3. Omit syllable | | | | | | | | | | | |
| 4. Insert syllable | | | | | | | | | | | |
| 5. Rearrange letters of syllables | | | | | | | | | | | |
| 6. Incorrect pronunciation { First | | | | | | | | | | | |
| { Last | | | | | | | | | | | |
| { Any Other | | | | | | | | | | | |
| a. Omit final sounds | | | | | | | | | | | |
| b. Shur final sounds | | | | | | | | | | | |
| c. Inarticulate vowels | | | | | | | | | | | |
| 4. Enunciation { d. Inaccurate vowels | | | | | | | | | | | |
| { e. Inarticulate consonants | | | | | | | | | | | |
| { f. Inaccurate consonants | | | | | | | | | | | |
| { g. Entire word indistinct | | | | | | | | | | | |
| 5. Substitutions { a. Meanings changed | | | | | | | | | | | |
| { b. Meanings unchanged | | | | | | | | | | | |
| 6. Insertions { a. Meanings changed | | | | | | | | | | | |
| { b. Meanings unchanged | | | | | | | | | | | |
| 7. Omissions { a. Meanings changed | | | | | | | | | | | |
| { b. Meanings unchanged | | | | | | | | | | | |
| 8. Other types of error | | | | | | | | | | | |
| II. Groups of Words | | | | | | | | | | | |
| 1. Change order { a. Meaning changed | | | | | | | | | | | |
| { b. Meaning unchanged | | | | | | | | | | | |
| 2. Add words to complete meaning according to fancy | | | | | | | | | | | |
| 3. Omit one or more lines | | | | | | | | | | | |
| 4. Insert two or more words { a. Meanings changed | | | | | | | | | | | |
| { b. Meanings unchanged | | | | | | | | | | | |
| 5. Omit two or more words { a. Meanings changed | | | | | | | | | | | |
| { b. Meanings unchanged | | | | | | | | | | | |
| 6. Substitute two or more words { a. Meanings changed | | | | | | | | | | | |
| { b. Meanings unchanged | | | | | | | | | | | |
| 7. Repeat two or more words { a. To correct error | | | | | | | | | | | |
| { b. To secure meaning better | | | | | | | | | | | |
| { c. To clear up uncertainty | | | | | | | | | | | |
| 8. Other types of error | | | | | | | | | | | |
| Pupil's test record { Rate | | | | | | | | | | | |
| { Errors | | | | | | | | | | | |
| Standard Scores for the Grade { Rate | | | | | | | | | | | |
| { Errors | | | | | | | | | | | |
| Date of Each Test | | | | | | | | | | | |

FIG. 11. Gray's Oral Reading Test, individual record sheet. (By permission of Public School Publishing Company, Bloomington, Ill.)

TABLE 7. DIAGNOSTIC READING TESTS

| Name of test | Grades | Time to give | Characteristics | Publisher |
|---|---|--|---|---|
| Durrell Analysis of Reading Difficulty | 1-6 | 50 min. | See discussion in chapter | World Book Company |
| Van Wageningen and Dyvorak Diagnostic Examination of Silent Reading Abilities | Division 1: 4-5 Division 2: 6-9 Division 3: 10-12 | Part I: 5 min. No time limit on Part II (45 min.) and Part III (60-90 min.) | Norms based on 30,000 urban and 15,000 rural children | Educational Test Bureau |
| Gray Oral Reading Tests. Sets 1,2,3,4. | 1-8 | Time varies. Depends on reader. Comparatively short. | See discussion in chapter. Norms for rate and accuracy | Public School Publishing Company |
| Ingraham Clark Diagnostic Reading Tests | Primary: 1-3 Intermediate: 4-8 | Part I: 30 min. (about). Part II: no time limit—"Turn to next test when 90% are through." | Two forms for each division. Form 1 and Form 2 in grade 2. .94. Each part has a reliability within grade of .87-.95. Primary: ability to recognize word forms and likeness, and differences among words, both visual and auditory stimuli. Intermediate: parts 1 and 2; reliabilities vary (.82-.95); words similar and their opposites; auditory visual recognition, sentences and paragraph meanings; relevant and irrelevant statements to be judged | California Test Bureau of South California, Book Depository |
| Ophthalmograph | Any grade | Varies | Photographs the number of eye fixations, refixations or regressive fixations, recognition span, rhythm, eye coordination, reading speed, and rhythm | American Optical Company |

and rare, which filled the old spelling books, the number now to be learned in the elementary school has been reduced to some three to four thousand with slight variations according to author. (Breed selects 3,481 words; Starch, 2,626; Horn, about 3,000; Washburn, 3,585; and Tidyman, 3,000 to 3,500.) This list has been arrived at through investigations of the following:

1. Children's compositions¹ as written by 1,050 children in grades 2 to 8.

2. Correspondence of adults as collected from 3,500 letters written by adults.²

3. Words appearing in newspapers.³

4. Words used by authors in the better magazines.⁴ The words were collected from the writings of 40 authors in 11 different magazines.

5. Thorndike's *Teachers Word Book* which contains the 10,000 words which occur most frequently in (a) the English classics, including the Bible, (b) children's literature, (c) newspapers, (d) correspondence, and (e) books about sewing, cooking, farming, and the trades.⁵

6. Horn's list,⁶ which contains 10,000 words selected from various types of adult correspondence. This study of Horn took into consideration all previous studies and thus has influenced greatly present spelling lists.

From these studies of the occurrence of words some inferences could be made. In the first place, over 95 per cent of ordinary running words are composed of a comparatively few words (about a thousand). These words occur again and again. In the second place, there were many words which children needed to spell which adults rarely used and vice versa. In short, there was no one-to-one agreement between needs of adults and the needs of children in spelling. In the third place, there was some lack of agreement between the lists of words drawn up from correspondence and those obtained from the inspection of English classics and other literature.

¹ Jones, W. Franklin, *Concrete Investigations of the Material of English Spelling with Conclusions Bearing on the Problems of Teaching Spelling*. University of South Dakota, 1913.

² Anderson, W. N., *Determination of Spelling Vocabulary Based upon Written Correspondence*, Studies in Education, Vol. II, University of Iowa, 1921.

³ Eldridge, R. C., *Six Thousand Common English Words*. Niagara Falls, N.Y., 1911.

⁴ Starch, Daniel, *Educational Psychology*, rev. ed. p. 38. New York: The Macmillan Company, 1927.

⁵ Thorndike, E. L., *The Teacher's Word Book*. New York: Bureau of Publications, Teachers College, Columbia University, 1921.

⁶ Horn, Ernest, *A Basic Writing Vocabulary*, Monographs in Education, First Series, No. 4, University of Iowa, 1926.

From such studies as those just described it became evident that the ordinary individual needed to spell correctly a small body of words used in communication. This common body of words would form the central core to be studied by all children in the elementary school. Each child, above all, should learn to spell those words which he normally used in his own writing. Moreover, each area of study must be responsible for teaching the spelling and meaning of the words which formed the special vocabulary of that discipline. Tests of spelling were necessarily constructed to test success in the correct spelling of the common core of spelling words.

OBJECTIVES IN TEACHING SPELLING

If we admit the criterion of social usefulness as adequate, the objectives in teaching spelling are:

1. To be able to spell correctly words occurring in ordinary communication. This would mean correctness when attention was directed to the expression of an idea rather than to the word being spelled.
2. To understand the meaning of the words which are being spelled. In a great many instances, such as in spelling "capital" or "capitol," "principal" or "principle," "there" or "their," understanding of meaning is necessary for correct spelling.
3. To attain a feeling of doubt of the correct spelling of some words. Under these circumstances the dictionary habit is absolutely necessary for correct spelling. The worst errors of all appear when a student goes blithely on misspelling words day after day and believing that his spelling is correct.
4. To develop in the student or pupil a desire to spell correctly which is so strong that he will be willing to go to considerable pains to assure correctness. Perhaps an ideal should be developed for the child which was derived from a variety of unfortunate occurrences accompanied by rather dire consequences when words were misspelled.
5. A technique or method for learning to spell either old words that are misspelled or new words needed in communication.

TESTS OF SPELLING

Ordinarily we become acutely aware of a child's failure in spelling when he makes a low score on words spelled in a test battery. Under such conditions it is not entirely certain that his spelling is poor because the sample is so small. For this latter purpose it is necessary to test his spelling with many more words selected from a list whose frequency and social value have been determined.

Survey Spelling Scales of Test Batteries

The spelling tests given as a part of an educational battery are usually composed of words carefully selected from available lists.

The Stanford Achievement Test,¹ for example, is composed of 100 words arranged in the order of their spelling difficulty. The first four words are "it," "and," "ten," and "old"; the last four, "cafeteria," "rabid," "contemporaries," and "dirigible." The spelling test for each grade starts and ends at defined positions. The second grade spells the first 40; the third grade, the first 50; and the fifth grade starts at the twenty-first word and goes through the seventieth word. All the grades, except the second, spell 50 words. The manner of giving is as follows: (1) the word is pronounced, (2) the word is presented in a sentence which determines its meaning, and (3) the word is pronounced again and only then spelled. Two illustrations are:

42. *shed*—We keep our coal in a wooden *shed*—*shed*.

43. *afraid*—Don't be *afraid*. This dog doesn't bite—*afraid*.

The Metropolitan Achievement Tests¹ uses 75 words in its spelling scale and arranges them in the order of their spelling difficulty. In the intermediate and advanced batteries the first four words of the list are "pan," "rest," "sweet," and "glad"; the last three are "deterrent," "chauffeur," and "adequate." As in the Stanford Achievement Tests, pupils of each grade start at a different place and end at a defined place. Examples are:

1. Grade 5 starts at the first word and spells through No. 50.

2. Grade 6 starts at the sixth word and spells through No. 55.

3. Grade 8 starts at the twenty-sixth word and spells through No. 75.

The method of presentation is also the same as that of the Stanford Achievement Test. Examples are:

26. *toward*—He turned from her to face *toward* me—*toward*.

27. *advertise*—It pays to *advertise*—*advertise*.

28. *happened*—He did not know what had *happened*—*happened*.

In the California Achievement Tests (formerly called the Progressive Achievement Tests) 30 words arranged from easy to hard constitute the spelling test in the battery. For example, in the advanced battery the 30 words begin with "grocery," "doubt," and "concert" and end with "souvenir," "inflammable," and "conscientious." The method utilized consists of first pronouncing the word, then presenting the word in a sentence to show its meaning, and finally pronouncing it again before it is spelled.

¹ Items by permission of World Book Company, Yonkers, N.Y.

Other achievement batteries usually contain spelling scales composed of lists of carefully selected words, either to be spelled after dictation and definition or else embedded in sentences which are dictated and copied. In tests of spelling suitable for the high school or college, sometimes the correct spelling of a word appears among three or four misspellings of the same word.

Separate Spelling Tests

Spelling tests unconnected with test batteries usually include many more words to be spelled. It is thus possible to select words of about equal difficulty and combine them into several sets or tests. It might even be desirable to have a test each month and to graph the improvement or lack of it which obtains from one month to the next. Most of the words which are necessary in ordinary communication could be included in these monthly tests. Three tests will be described here: (1) the Ayres Spelling Tests, (2) the Iowa Spelling Tests (Ashbaugh), and (3) the Morrison-McCall Spelling Scale.

The Ayres Spelling Scale consists of 1,000 words most frequently used in written discourse. They were selected from 368,000 running words written by 2,500 different persons. These words were selected from a list which was combined from four studies of words used in newspapers, good literature, and letters. The difficulty of the words was determined by submitting 50 lists of 20 words each to children to be spelled. Two consecutive grades of children spelled each list. Altogether these 1,000 words were spelled by some 70,000 grade school children living in 84 cities in different parts of the United States.¹ An average of 1,400 spellings was made of each word. In this manner, the difficulty of each word for each grade was determined. Words of similar difficulty are arranged in 26 columns from A to Z. The scale consists of this 1,000 most useful words arranged on a single sheet with small numbers of words at the ends and many toward the middle. Under each letter and just above the list of words are a set of percentages which indicate the percentages correct which were spelled by certain grades. For example, under the letter O are the percentages 27, 50, 73, 84, 92, 96, and 99, which are an indication of the words correctly spelled by grades 2 to 8 respectively. In preparing a test from this list of 1,000 graded words, the best procedure is to select about 25 words from the column where about 50 per cent of correct spelling is anticipated. If the class varies greatly in its ability to spell, or if there are not sufficient words in the appropriate column, some of the words may be selected from the less difficult and some from the more difficult

¹ Ayres, L. P., *Measurement of Ability in Spelling*, Bulletin of the Division of Education, New York.: Russell Sage Foundation, 1915.

columns. It is permissible to use words varying in difficulty from 16 per cent to 84 per cent (*i.e.*, \pm one S.D. from the mean) expectancy, since such a procedure tends to distribute the pupils' spelling scores on a normal curve.

As now printed, the Ayres Spelling Scale becomes the Buckingham Extension of the Ayres Spelling Scale. Buckingham added 505 words to Ayres's 1,000.

The Iowa Spelling Scales, published in 1919, use 2,997 words instead of the 1,000 used by Ayres. These words were found by Anderson to be most frequently used in the written correspondence of adults.¹ Ashbaugh,² the author of these scales, arranged the words in seven scales intended for use in grades 2 to 8. The difficulty of the words was determined by 200 spellings of each word of the scale. The words constituting each test and suitable for a certain grade are arranged in groups according to difficulty. Here is an example for grade 5 with different percentages of words spelled in the standardization of the test.

| 53 per cent | 54 per cent | 55 per cent | 59 per cent |
|-------------|-------------|-------------|--------------|
| advertised | affair | accept | alfalfa |
| article | awful | advancement | channel |
| assist | considered | advertise | connected |
| automatic | corrected | agreeable | contemplated |
| carrying | correction | attended | decided |
| (out of 23) | (out of 29) | (out of 26) | (out of 31) |

You will note that the words within the grade vary slightly in difficulty. For the test proper the words which constitute the test should be selected from those whose difficulty approximates 50 per cent. Better results are obtained when the difficulty approaches 50 per cent because such difficulty offers opportunity for measuring adequately both the poor and the excellent spellers.

The Iowa Spelling Scale has certain advantages of great importance:

1. It contains 2,997 well-graded words already arranged in order of spelling difficulty.

2. The words are socially highly useful.

3. The words in the test can be used as a criterion of social usefulness against which to project words in the ordinary spelling book.

The Morrison-McCall Spelling Scale consists of eight lists of words together with the illustrative sentences. The 50 words in each list are suitable for testing the spelling ability of children in grades 2 to 8. The manner of presenting the words is illustrated from List 1:

¹ Anderson, *op. cit.*

² Ashbaugh, E. J., *The Iowa Spelling Scales*. Bloomington, Ill. Public School Publishing Company, 1922.

15. *done*—Has he *done* the work?—*done*.
 39. *reference*—He made *reference* to the lesson—*reference*.

In each list the words are arranged from easy to hard but the eight lists are equal in difficulty. "All the words in each list of this spelling scale were selected from Ayres' Spelling Scale and Buckingham's Extension of Ayres' Spelling Scale, in such a way as to make all lists equally difficult, and the words were required in addition to appear among the 5,000 most commonly used words as reported in Thorndike's *Word Book*."¹

Norms are furnished for each grade from 2 to 9 as well as for each age. It is clear that this scale furnishes a well-defined procedure for administering words from the Ayres Scale.

OTHER TESTS OF SPELLING

1. Public School Achievement Test in Spelling, grades 2-8. Four forms. All test words from the Iowa Spelling Scales. Public School Publishing Company, Bloomington, Ill.

2. Courtis Standard Research Tests in Spelling, grades 2-8. S. H. Courtis, Detroit, Mich.

3. Davis-Schrammel Spelling Test,

all grades. Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kans.

4. Unit Scales of Attainment in Spelling, grades 3-8. Educational Test Bureau, Minneapolis, Minn.

5. The Gates-Russell Spelling Diagnosis Tests, 1937.²

Uses of Spelling Scales and Tests

First and foremost, spelling tests can be used to make certain that the children can spell the 3,000 words so necessary for ordinary communication as well as the average of the country. Do the children spell correctly as many words as the norms demand? If they do, some teachers and administrators are then satisfied. But this is not enough. If the 2,997 words of the Iowa Spelling Scales represent the minimum essentials in spelling, then the vast majority of children should be able to spell all the words. It is here that such scales have their greatest use.

Spelling scales, since they are easy to use and to score, can be given frequently and the results graphed not for public consumption but to study each individual case. Thus a child can be easily taught to make a bar graph indicating the number of words he has spelled correctly (1) on the first test, (2) on the second test, (3) on the third test, etc. He has the evidence clear and unmistakable as to his progress in spelling.

¹ *Manual*, p. 7. Yonkers, N.Y.: World Book Company, 1923.

² Gates, A. I., and D. H. Russell, *Diagnostic and Remedial Spelling Manual*. New York: Bureau of Publications, Teachers College, Columbia University, 1937.

He should also make a list of the words he has spelled incorrectly in each of the tests. *Such a procedure constitutes a genuine motivating force.*

The third use of spelling scales arises in connection with analyzing the spelling difficulties which each child encounters. After such a spelling test has been taken the misspelled words are collected and a study made of them to discover if there are any recurrent errors. Are the words incorrectly spelled because certain consonants were not doubled, certain connecting vowels were interchanged, or there was confusion between "-ance" and "-ence," etc.? Maybe some rules of spelling will aid here. In any case the student now knows what part of the word to study more attentively.

From such spelling tests there are found students whose spelling is poor indeed. Gates and Russell tells of a pupil who was able to spell only 7 out of 55 words dictated.¹ Such cases need a detailed analysis of their spelling difficulties. All data that might possibly influence their spelling deficiencies should first be collected. The records of their reading and spelling tests and possibly scores on handedness or eyedness tests must be brought together. If such scores do not exist tests should be immediately given and the scores assembled. The inspection of such data may give an immediate insight into the general difficulties. Furthermore, investigation must be made of all other traits which bear or might bear upon the problem. Tests of visual and auditory discrimination are administered as well as tests of intelligence. Checks are made on the individual's *visual or auditory memory* which may be so poor that he cannot remember how a word looks or sounds long enough to write it down correctly. His pronunciation is checked, for it may be so faulty that letters are omitted or wrongly entered. If a child pronounces "courthouse" "c-o-a-thouse" or "bird" "b-o-i-d" his spelling difficulty is increased. Perhaps the pronunciation of "February" is a more universal example. But for detailed analysis further diagnoses are needed.

Fortunately, we have such a test or series of tests, Gates-Russell Spelling Diagnosis Tests,² which attempt to discover why children use reversals, insertions, omissions, substitutions, transpositions, phonetic errors, additions, etc., in their spelling. More precisely stated, this test furnishes a method of discovering the errors in the nine areas listed below. From their investigations and their psychological insight they developed this series of diagnostic tests, which they list as follows:

1. Spelling words orally
2. Word pronunciation
3. Giving letters for letter sounds
4. Spelling one syllable (nonsense syllables)

¹ *Ibid.*, pp. 30-31.

² *Ibid.*

5. Spelling two syllables (nonsense syllables)
6. Word reversals
7. Spelling attack—method of study
8. Auditory discrimination
9. Visual, auditory, kinesthetic, and combined study methods

By means of these tests, which a teacher, with a little practice, can easily learn to give, analysis can be made of the sources of error and learning and practice directed to the areas where it will count most. In some cases, it is discovered that the pupil never has learned a good method of studying a word. For this reason, he must be taught how to *learn* to spell.

HANDWRITING

As a means of communicating with others, handwriting has lost some ground in the last fifty years to typewriters and other sorts of recording machines. However, in social communication, and in making private notes it still maintains an important position. It is also the principal means of helping the student clarify his own thoughts on any topic. "Writing maketh an exact man."

The genetic development of this complicated motor habit throws some light on its complexity. In the early stages of learning, handwriting is largely a matter of perceptual motor learning. The child looks at the letter and then draws it. Tracing it by means of overlaid tissue paper or controlling the direction of movement by means of holding the child's hand is of little value. He must learn to draw what he sees. The model is before him. As time goes on, the perceptual object or model is removed and learning becomes ideomotor. These simple, disparate habits must be integrated in such a way that the handwriting is smooth and rapid. Sometimes smoothness and speed play hob with the letter forms, so that improvement in those respects is gained at the expense of quality. Every so often the learner must return to improving the quality in more or less formal exercises.

Probably the greatest enemy of quality, therefore, is the shift of attention from form to substance as one writes. If one's thoughts shift to form and legibility, they improve but ideas and organization suffer. For this reason, quality and speed of handwriting must be learned so well that they are almost self-running.

AIMS AND OBJECTIVES IN TEACHING HANDWRITING

The aims and objectives of the teaching of handwriting are determined by levels of attainment achieved by pupils in the appropriate grades. In the earlier grades norms for speed and quality are defined by those levels of attainment that children under good instruction

have succeeded in reaching. But in the upper grades adult standards are the determining factors. Questions of how well employers expect employees to write without being penalized in their work enter into the norms for achievement.¹

One investigator (Koos) studied the quality of 1,053 specimens secured from social correspondence and 1,127 samples of the handwriting of employees from a large variety of occupations. Furthermore, he obtained from 826 adults their opinion as to what was satisfactory or unsatisfactory for social correspondence. When the results of what children accomplish and employers desire are compared, the conclusion arrived at is that children should be taught to *write as well as quality 60 on the Ayres Handwriting Scale and at the rate of 70 letters a minute*. Since speed is closely correlated with age, the rate in high school could be pushed up to 80 or 90 letters per minute without greatly affecting the quality.

A second aim is to teach pupils how to analyze their own difficulties in handwriting and to instruct them in a method by which their improvement will be assured. Along with this aim is the developing of the attitude in children that good body posture is helpful especially if much writing is to be done.

The third objective in the teaching of handwriting is to teach pupils to place their writing properly on a page. Here instruction in the use of headings, margins, and spacing seems most important. You will see that one of the variables on the Freeman Handwriting Chart is spacing.

A fourth aim is to teach pupils to want to write well whenever handwriting is to be done. Slovenly habits in handwriting are due pretty largely to a failure of the pupil to realize the importance of writing well.

In short, a child who writes well enough for satisfactory social communication, who has learned a method of analyzing and improving his own handwriting, who arranges the written material properly on a page, and who desires to write well at all times has fulfilled the objectives of handwriting in the elementary school.

MEASUREMENT OF HANDWRITING

Both the rate and quality of handwriting have been measured. These two variables are interdependent. To a very considerable extent they depend upon the "set" of the subject. If the set is obtained by instructing the subject to "write as rapidly as possible," then quality suffers. If the set is obtained by asking the subject to "write as well as

¹ Koos, L. V., "The Determination of Ultimate Standards of Quality in Handwriting for the Public Schools," *Elementary School Journal* (1918) 18:422

possible," rate suffers. For example, one experimenter (Freeman, 1915) showed that when he called for quality, speed was reduced 3.7 per cent and quality improved 6.2 per cent. On the other hand, when he called for speed, quality decreased 9.1 per cent but rate increased 27.2 per cent. In general, the instructions to subjects to obtain the best results should be, "Write as well as you can and as rapidly as you can."

Rate

It is comparatively easy to secure reliable measures of rate provided a few simple precautions are taken. When a subject is being measured for simple rate of handwriting we must not confuse the issue by introducing other variables. Satisfactory results are obtained if the child knows the material by heart, if he can easily spell all the words, and if the words are not too long. It is customary to use the same sample for both rate and quality. In this case, the material should be the same as that appearing in the rating scale.

If one were administering the Gettysburg edition of the Ayres Scale, he would copy on the board Lincoln's Gettysburg Address. He would then go over it with the children, calling attention to the words and their spelling. When the subjects are thoroughly acquainted with the passage they would secure pen and ink and copy the material as well and as rapidly as they could. Two minutes are used for writing. Scoring is facilitated by putting after each word on the scoring sheet a number indicating the total number of letters written to that point. For example: "Four 4 score 9 and 12 seven 17 years 22," etc. Speed of writing is the number of letters written per minute.

There is considerable justification for using simpler, more interesting material for the lower grades. Thus the American Handwriting Scale¹ uses words chosen with the child's interests in mind: "Anna 4 has 7 six 10 dear 14 baby 18 kittens 25. They 4 play 8 with 12 a 13 round 18 red 21 ball 25." The units are 25 letters long. The reliabilities of rate scores are high indeed.

Quality

In measuring quality it is necessary to compare the subject's sample of handwriting with a set of samples whose qualities, ranging from poor to excellent, have already been determined. Such sets of graduated samples are (1) the Thorndike Scale for Handwriting of Children, (2) the Ayres Measuring Scale for Handwriting (Gettysburg edition) and (3) the Conard Manuscript Writing Standards.

¹ West, Paul V., *American Handwriting Scale*, grades 2-8. Chicago: A. N. Palmer Co., 1929.

The Thorndike Scale for Handwriting of Children¹ was first published in 1910. It was the first scientifically constructed instrument for educational measurement. This instrument consists of samples arranged along a scale from No. 4, which was artificially constructed, to No. 18, which is a copybook model. All the rest of the samples were written by children. The differences between samples were determined by competent judges who voted one sample to be better or worse than another on the basis of *general merit*. If 75 per cent of equally competent judges judged one sample to be better than another, then the sample so judged was taken as one unit better than the other. For example, if 75 per cent said sample 7 is better than sample 6 then sample 7 is one unit (probable error) above sample 6. Thus the scale was constructed so that each succeeding sample was one unit better than each preceding. We thus have a scale—4, 5, 6, 7, etc.—in which the differences between units up and down the scale are approximately the same.

Some weaknesses have caused the Thorndike scale to be less used than formerly. The samples of the scale do not resemble closely enough the type of handwriting now prevalent. Nor do all the samples contain copy of exactly the same material. This makes it more difficult to compare with a new sample. The number of samples appearing at each scale unit varies from one at numbers 4, 5, 6, and 7, two at 8, three at 9, two at 10, to four at 15. The norms that appear on the scale are shown in the accompanying table.

HANDWRITING STANDARDS

| | Grade | | | | | | |
|--|-------|-----|------|------|------|------|------|
| | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Speed, letters per minute. . . . | 35 | 45 | 55 | 64 | 72 | 77 | 80 |
| Quality as measured on the Thorndike scale: | | | | | | | |
| Usual. | 7.0 | 7.8 | 8.6 | 9.3 | 9.9 | 10.5 | 11.0 |
| Best | 8.5 | 9.3 | 10.1 | 10.8 | 11.4 | 12.0 | 12.5 |

The Ayres Measuring Scale for Handwriting² differs from the Thorndike scale in several particulars. In the first place the basis for judgment of position in the scale is *legibility*. Legibility was thought to be more

¹ New York: Bureau of Publications, Teachers College, Columbia University.

² Bloomington, Ill.: Public School Publishing Company.

functional and more objective than "general merit." The samples were read by 10 paid assistants who kept careful records of the time of each reading. The scale value of each sample was determined by the average time used up by the 10 assistants in its reading. The scale consists of eight samples written in blue ink and ranging in value from 20 to 90. Some critics have felt the need of a score lower than 20 and of one above 90, but one can interpolate a 15 or a 95 without making too great an error.

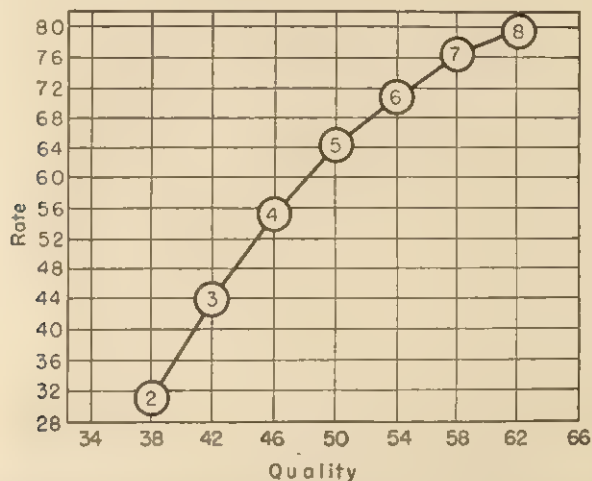


FIG. 12. Norms, Ayres Handwriting Scale. (By permission of Department of Education, Russell Sage Foundation, New York.)

The Ayres Measuring Scale for Handwriting, Gettysburg edition, has become the most used of all the handwriting scales. The norms are given in Fig. 12.

The Conard Manuscript Writing Standards¹ was developed to meet the need of large numbers of teachers of the primary grades who introduce their pupils to writing by using the manuscript method. There are two sets of scales: (1) for pencil, and (2) for pen. The pencil scale is composed of samples 1 to 12 selected from 5,000 samples of manuscript writing. The samples vary in quality from No. 1, which is practically illegible, to No. 12, which is quite satisfactory for students in Grade 11 (Fig. 13). This scale is suitable for grades 1 to 4. The scale for pen has 10 samples, reaching from third grade to adult level. The last two samples were written by adults. All the rest were written by children in grade 6 or below.

¹ New York: Bureau of Publications, Teachers College, Columbia University. Items by permission.

3

we made
apple jelly.

6

Dear Miss Conard
I am glad that we can
our writing.

10

Sunday, Monday, Tuesday. We
Thursday, Friday, Saturday.
Sunday, Monday, Tuesday. W
Thursday, Friday, Saturday,

FIG. 13. Conard Manuscript Writing Standards, Samples 3, 6, and 10 (pencil).

Use of Quality Scales

The sample which you wish to rate is slid along until it matches a sample on the scale, say 50 on the Ayres scale. This 50 is put on the back of the sample. The papers are now shuffled and the same process is repeated. It is thus possible to obtain two independent scores, a fact which makes for accuracy. The score on the second rating is then averaged with the score on the first. This average constitutes the score for the paper. Best results of all are obtained by having each paper rated independently by three different persons.

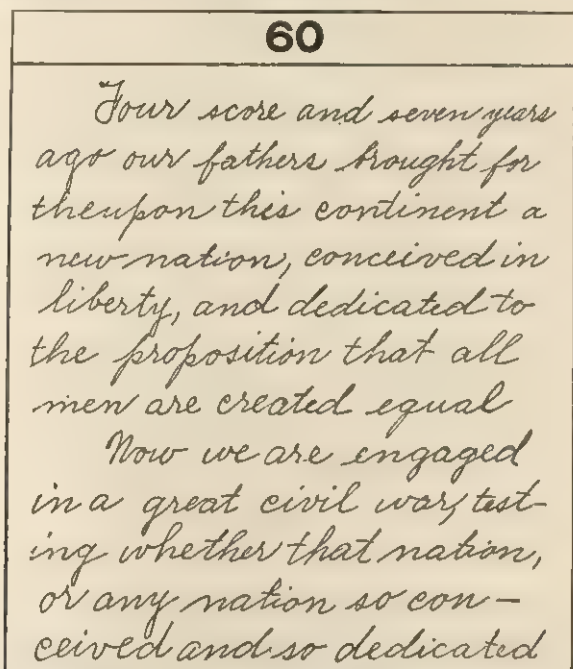


FIG. 14. Sample 60, Ayres Handwriting Scale. (By permission of Department of Education, Russell Sage Foundation, New York.)

For those teachers truly concerned about the reliability of quality ratings, practice in rating Thorndike's 50 samples is recommended.¹ In this case each sample is scored independently and then compared with the score agreed upon by experts (the true score). By means of such practice considerable gains in accuracy can be achieved.

Reasonable Quality to Be Expected

It is the opinion of experts in the field, of employers, and of a large number of the general run of people that quality 60 on the Ayres scale

¹ Thorndike, E. L., "Teachers' Estimates of Specimens of Handwriting," *Teachers College Record* (November, 1914) Vol. 15, No. 5.

written at the rate of about 70 letters per minute is a reasonable achievement in handwriting at the end of grade 6. Figure 14 shows sample 60 of the Ayres Handwriting Scale.

DIAGNOSIS AND ANALYSIS OF HANDWRITING

Progress in motor learning begins with a general understanding of the problem, proceeds by means of trial and error, with some limita

Letter Formation

A quick brown fox jumps over the lazy dog.

A quick brown fox

jumped over the lazy

A quick brown fox jumps over the lazy dog

A quick brown fox jumps over

A quick brown fox jumps over the lazy dog

A quick brown fox jumps over the lazy dog

Some books are to be tasted; others to be swallowed, and some few to be

FIG. 15. Freeman's Chart for Diagnosing Faults in Handwriting, letter formation. (By permission of Houghton Mifflin Company, Boston.)

tion of error through guidance and practice, and results in the establishment of a certain level of achievement. If this level of achievement is low, further progress is contingent upon the analysis of habits and the direction of practice toward a much narrower function. Such analysis of habits may take place in handwriting. Three procedures

aimed at diagnosis and improvement will be mentioned: (1) Freeman's Chart for Diagnosing Faults in Handwriting,¹ (2) Gray's Score Card for Measuring Handwriting,² and (3) Freeman's Score Card of Defects in Handwriting.³

Freeman's Chart for Diagnosing Faults in Handwriting divides handwriting into five separate traits: (1) uniformity of slant, (2) uniformity of alignment, (3) quality of line, (4) letter formation, and (5) spacing. Each part or division appears on the sheet at three levels of performance, which have the accompanying scores: (1) poor, a score of 1, (2) average, a score of 3, and (3) good, or excellent, a score of 5. These five attributes of handwriting appearing at three levels of performance are printed on one large page. Uniformity of slant is judged by drawing lines parallel and close to the long letters such as h, t, or b. The scoring of uniformity of alignment is facilitated by drawing parallel lines above and below the written line. A reading glass aids in judging the quality of line. The diagnosing of correct letter formation is aided by a large number of little arrows pointing to poorly formed parts of letters. A sample of that part of the scale called Letter Formation is shown in Fig. 15. Freeman recommends that one attribute be scored at a time, each independently of the others. By assuming the scores of an individual in each of the five traits a total rank is obtained. Such a chart, while not completely diagnostic, does tend to focus the teacher's thoughts on special aspects of handwriting which need improving. It may also suggest that if one attribute is practiced at a time greater improvement in handwriting may be attained.

Score Cards

Score cards attempt to describe in words and to arrange in a sort of check list the elements which compose handwriting. Their best use is in diagnosing difficulties rather than in giving a total score or rating.

Gray's Standard Score Card for Measuring Handwriting, which appears in Fig. 16, not only lists the qualities to be studied but weights them so that the total points of a perfect handwriting would be 100. The formation of letters, so essential to legibility, is given a score of 26, the largest weight.

Freeman's Check List of Defects in Handwriting, which is to be used in connection with his Chart for Diagnosing Faults in Handwriting, not only lists the defect but describes its most probable cause:

¹ Boston: Houghton Mifflin Company.

² Bloomington, Ill.: Public School Publishing Company.

³ Freeman, F. N., *The Teaching of Handwriting*. Boston: Houghton Mifflin Company, 1914.

Defect

1. Too much slant (1) Writing arm too near body
 (2) Thumb too stiff
 (3) Point of nib too far from fingers
 (4) Paper in wrong direction
 (5) Stroke in wrong direction

West's Score Sheet for Diagnosis of Defects in Samples of Handwriting¹ and the Pressey Chart for Diagnosis of Illegibilities in Handwriting¹ are other instruments used for diagnosis of handwriting. Such check lists can be of some help to the teacher of handwriting. Probably actual visual illustrations such as appear in the diagnostic charts give more effective help in diagnosing difficulties of handwriting.

Practice Exercises

There are three sets of practice exercises which warrant study here. They are (1) Courtis-Shaw Practice Tests in Handwriting,² (2) Leamer Diagnostic Practice Tests in Handwriting,¹ and (3) Minneapolis Self-Correction Handwriting Charts.³ These three have certain characteristics in common. All three provide opportunities for discovering defects in writing and offer some opportunity for improving the conditions found. All assume that each child will practice on his own difficulties and proceed at his own rate. The Courtis-Shaw and Leamer instruments provide for graded standards of accomplishment. The attainment of these standards may enable the pupil, if he is in the lower grades, to go on to more difficult writing tasks or, if he is in the upper grades, may excuse him from further practice.

In the Leamer Diagnostic Practice Sentences in Handwriting the pupil practices a simple sentence for 8 minutes. He then writes for 2 minutes what he has practiced. Before him is a sample which is the standard for his grade. As soon as he writes the sentence which he has practiced as well as the standard and at the proper rate, he is then given another sentence to practice. Each child has an opportunity to measure his own success or failure. The Courtis-Shaw Standard Practice Tests in Handwriting require the subject to take a preliminary handwriting test which is intended to reveal his present proficiency in this subject. The Minneapolis Self Correcting Handwriting Charts, devised by Nystrom, provide for a very thoroughgoing diagnosis of handwriting defects. On one side of the chart are opportunities for analyzing defects such as letter and word spacing, alignment, slant,

¹ Bloomington, Ill.: Public School Publishing Company.

² Yonkers, N.Y.: World Book Company.

³ St. Paul, Minn.: St. Paul Book and Stationery Co.

Standard Score Card for Measuring Handwriting

By
C. TRUMAN GRAY

Pupil Age Date

Grade School

Sample Number Teacher

| Sample | Perfect Score | 1st mo. | 2nd mo. | 3rd mo. | 4th mo. | 5th mo. | 6th mo. | 7th mo. | 8th mo. | 9th mo. | 10th mo. |
|-------------------------------|---------------|---------|---------|---------|---------|---------|---------|---------|---------|---------|----------|
| 1. Heaviness | 3 | | | | | | | | | | |
| 2. Slant | 5 | | | | | | | | | | |
| Uniformity | | | | | | | | | | | |
| Mixed | | | | | | | | | | | |
| 3. Size | 7 | | | | | | | | | | |
| Uniformity | | | | | | | | | | | |
| Too large | | | | | | | | | | | |
| Too small | | | | | | | | | | | |
| 4. Alignment | 8 | | | | | | | | | | |
| 5. Spacing of lines | 9 | | | | | | | | | | |
| Uniformity | | | | | | | | | | | |
| Too close | | | | | | | | | | | |
| Too far apart | | | | | | | | | | | |
| 6. Spacing of words | 11 | | | | | | | | | | |
| Uniformity | | | | | | | | | | | |
| Too close | | | | | | | | | | | |
| Too far apart | | | | | | | | | | | |
| 7. Spacing of letters | 16 | | | | | | | | | | |
| Uniformity | | | | | | | | | | | |
| Too close | | | | | | | | | | | |
| Too far apart | | | | | | | | | | | |
| 8. Neatness | 13 | | | | | | | | | | |
| Blotches | | | | | | | | | | | |
| Carelessness | | | | | | | | | | | |
| 9. Formation of letters | (26) | | | | | | | | | | |
| General form | 8 | | | | | | | | | | |
| Smoothness | 6 | | | | | | | | | | |
| Letters not closed... | 5 | | | | | | | | | | |
| Parts omitted | 5 | | | | | | | | | | |
| Parts added | 2 | | | | | | | | | | |
| TOTAL SCORE | | | | | | | | | | | |

Scored by

Distributed by the Public School Publishing Co. of Bloomington, Ill.

111-1p

FIG. 16. Gray's Standard Score Card for Measuring Handwriting. (By permission.)

etc. On the other side of the chart are exercises for correcting these defects.

Uses of Scales and Check Lists in Improving Handwriting

From the standpoint of the administrator there are several uses of measurement of rate and quality in handwriting. Comparisons may be made between the school grades of his system and the assembled norms as well as between different school grades. Through such procedures he can obtain an over-all picture of the pupil progress in handwriting in a single school or in his total school system.

The teacher finds in the scales and in the norms for quality and rate of handwriting attainable goals of achievement. To develop in pupils the ability to write as well as 60 on the Ayres scale at a rate of 70 letters per minute defines objectively the aims of instruction. He can *know* what is expected of pupils in this school subject. Greatest help of all, however, comes to the teacher in the form of diagnostic scales such as Freeman's or in check lists of possible defects. These instruments define and narrow the problems of instruction but do not solve them. To be solved there are needed practice exercises by means of which pupils may practice at the point of error. Under such conditions rapid improvement can be made in handwriting.

The pupil himself may find these scales and charts of handwriting of great value. If the Ayres scale is pasted on a planed pine board and hung at an easily accessible level, pupils may estimate directly their own progress. Such contemplation of attainable goals is a stimulating and motivating influence. If a child is taught to use the diagnostic charts and the check lists their value is greatly increased. Instead of practicing handwriting by "aiming at it generally," he learns to practice the letter or the parts of letters which are poorly formed. His rate of progress, too, depends upon himself—a condition which is in itself a motivating influence.

It thus becomes clear that handwriting scales and check lists may be used both to improve learning and to aid in evaluating the handwriting products which the subjects have produced.

SUMMARY

Tests of reading, spelling, and handwriting are described in this chapter.

Reading tests in the areas of reading readiness, reading achievement, and reading diagnosis are presented. Tests of reading readiness sample those mental processes which are usually deemed necessary for beginning the more formal aspects of reading. The levels of achievement in

vocabulary, language usage, knowledge of ordinary facts and events, number information, etc., are tested. Achievement is measured by tests which ask questions about vocabulary, the meaning of sentences and paragraphs and sometimes about the uses of indexes for locating information. The tests attempt to parallel teaching procedures with their questions and to test for the fulfillment of the objectives for which the teachers are striving. Diagnostic tests attempt to discover those sources of weakness which prevent the comprehension of the meaning of what is read.

Perhaps in no other subject are the aims and objectives more adequately defined than in spelling. The criterion of curricular validity is satisfied because the words to be tested are fairly well agreed upon. The words, about 3,000 in number, were arrived at through analysis of correspondence, articles, and books which well-known authors had written, and through a consideration of children's literature, newspapers, and the English classics, including the Bible. From these sources words needed by every one for ordinary correspondence were collected and graded for difficulty. Spelling scales and tests have been constructed using these very words. The aims of teaching spelling are for the most part realized when children can spell these words in addition to the words used in their ordinary communication. Such tests have been shown to be usable in (1) checking the general spelling level of a class, (2) motivating the individual's spelling procedures, and (3) analyzing the individual's misspelled words so that his accuracy in spelling might be greatly increased. It was also suggested that each pupil be taught a satisfactory technique for learning thoroughly the correct spelling of a word. Achievement in handwriting is dependent upon rate and quality. It is easy to measure rate, for it is simply the number of letters written per minute with material which is well known to the subject and whose spelling difficulties have been removed. Quality is estimated by comparing the subject's sample of writing with a series of samples whose positions in the quality scale have been previously determined. The second type of scale divided handwriting into five elements: slant, alignment, quality of line, letter formation, and spacing. Each of these five elements was presented at three levels: poor, average, and good. Check lists composed of words descriptive of defects in writing were also introduced. The hope was to assist teachers and pupils to discover exact points in handwriting where there were defects. To obtain most satisfactory results from such diagnoses, practice exercises must be provided so that effective practice may be directed toward those points which diagnosis showed were in greatest need of improvement. The Leamer and Courtis-Shaw are examples of practice exercises with clear-cut standards of achievement before each

learner. Practice thus becomes an individual matter, with each child progressing at his own rate.

In these three areas of measurement of the language arts there seems to be a definite need for a general survey test followed by tests of diagnosis. These needs parallel the purposes for which tests are used. The survey tests are of most use to the administrator, who wishes to know the present status and progress of grades as a whole and in various school buildings. The diagnostic tests are of greatest use to the teacher and the pupil. The teacher, interested in pupil improvement, discovers from these tests points where practice counts most. The pupil, too, may be stimulated to undertake a particular activity when the goal of total improvement would seem too distant.

QUESTIONS AND EXERCISES

I. READING

1. Name and explain five or six developmental traits on which reading readiness depends.

2. In what respects are reading-readiness tests different from intelligence tests in predicting reading readiness?

3. Secure a copy of the Lee-Clark Test of Reading Readiness and compare point by point with the Metropolitan. Which is superior for the purpose at hand? Evidence?

4. Describe the leading characteristics of the Iowa Silent Reading Test. How does it differ in construction from the Gates Silent Reading Tests? Which one do you think more nearly parallels a normal situation in reading? Why?

5. Describe in detail the procedures used to discover the causes of poor reading. Do you think the word "diagnostic" a proper one for describing what the test does?

6. How are tests of oral reading sometimes useful in discovering errors which affect the efficiency of silent reading?

7. Plan out a schedule of testing for studying the problem of reading in the ordinary school.

II. SPELLING

1. *a.* Describe the various procedures used in deciding upon the words whose spelling was to be learned in the elementary school.

b. What are the aims of the school in spelling instruction?

2. *a.* What limitations appear in the spelling test in the usual test battery?

b. How can the separate test get rid of these limitations?

c. What method of presentation of words to be spelled is common to many test batteries?

3. Describe and illustrate the uses to which spelling tests may be put.

4. What advantages do you see in such a test as the Gates-Russell Spelling Diagnosis Tests over the usual battery?

5. Explain why many students regard the Iowa Spelling Scales as the best of all spelling tests.

6. What is a good method for students or pupils to use in learning to spell new words? Explain in detail.

7. What range of difficulty would you use in constructing a test from the Ayres scale? Why?

8. Describe the process you would use in checking the social usefulness of the words included in a spelling book.

III. HANDWRITING

1. Secure 15 or 20 samples of children's writing of as much of the Gettysburg Address as they can get finished in 2 minutes.

a. Rate the quality on the Ayres scale, on the Thorndike scale. Place the scale value on the back of each paper.

b. After the papers have been shuffled, rate them again a second time, placing the score in front. Average the two marks. If there is a wide discrepancy between the two scores in the case of any paper, rate it again a third time.

c. Erase your marks and get another member of your college class to rate them. The average mark for any one paper is probably the best indication of its true position on the scale.

2. Secure a copy of Freeman's diagnostic chart.

a. Rate the papers on each element of the chart.

b. Combine the ratings of the five elements.

c. How does this total agree with the rating of the same paper on the Ayres scale?

d. Analyze the difficulties of several papers and test the defects on each paper.

e. Apply the Gray Score Card to the same paper.

3. a. Compare the efficiency of the Thorndike and Ayres scales in measuring samples of handwriting.

b. Which is most useful? Why?

4. Which do you think is the better procedure to get a satisfactory judgment about the quality of handwriting: (a) by the use of a general scale such as the Ayres, or (b) by scoring the paper on five elements and summing these scores? Explain.

5. What is the relation between handwriting rate and age?

6. Describe the leading characteristics of standardized practice exercises in handwriting. Illustrate by referring to the specific exercises described in the text.

7. To what uses can the instruments for measuring the quality of handwriting be put by (a) the administrator, (b) the teacher, (c) the student?

BIBLIOGRAPHY

I. READING

Books

BETTS, E. A.: *The Prevention and Correction of Reading Difficulties*. Evanston, Ill.: Row, Peterson & Company, 1936.

DURRELL, DONALD D.: *Improvement of Basic Reading Abilities*. Yonkers, N.Y.: World Book Company, 1940.

GATES, A. I.: *The Improvement of Reading: A Program of Diagnostic and Remedial Methods*, 3d ed. New York: The Macmillan Company, 1947.

———: *Methods of Determining Reading Readiness*. New York: Bureau of Publications, Teachers College, Columbia University, 1939.

GREENE, HARRY A., ALBERT N. JORGENSEN, and J. RAYMOND GERBERICH: *Measurement and Evaluation in the Elementary School*, "Reading," Chap. XV, "Spelling," pp. 373-384; "Handwriting," pp. 384-399. New York: Longmans, Green & Co., Inc., 1942.

HARRISON, M. LUCILE: *Reading*

Readiness. Boston: Houghton Mifflin Company, 1936.

MONROE, MARION: *Children Who Cannot Read*. Chicago: University of Chicago Press, 1932.

——— et al.: *Remedial Reading*. Boston: Houghton Mifflin Company, 1937.

TIEGS, ERNEST W.: *Tests and Measurements in the Improvement of Learning*, pp. 110-125, 159-165. Boston: Houghton Mifflin Company, 1939.

TOWNSEND, AGATHA: "A Study of the Revised New Edition of the Iowa Silent Reading Tests," pp. 31-39, in 1944 *Fall Testing Program in Independent Schools and Supplementary Studies*, Educational Records Bulletin. New York: Educational Records Bureau, 1945.

WEBB, L. W., and ANNA MARKET SHOTWELL: *Testing in the Elementary School*, Chaps. VIII-IX, "Spelling," Chap. 11, "Handwriting," pp. 231-254. New York: Rinehart & Company, Inc., 1939.

Articles

GRAY, WILLIAM S.: "Reading," *Encyclopedia of Educational Research* (Walter S. Monroe, ed.), pp. 891-926. Also "Reading—II. Physiology and Psychology of Reading," rev. ed., pp. 972-1005. New York: The Macmillan Company, 1941 and 1950.

KILBY, RICHARD W.: "Relation of Iowa Silent Reading Test Scores to Measures of Scholastic Aptitude and Achievement," *Journal of Applied Psychology* (1946) 30:399-405.

LEE, J. MURRAY, WILLIS W. CLARK, and DORIS MAY LEE: "Measuring Reading Readiness," *Elementary School Journal* (1934) 34:656-666.

STONE, CLARENCE R.: "Validity of Tests in Beginning Reading," *Elementary School Journal* (1943) 43:361-365.

WITTY, PAUL A., and DAVID KOPEL: "Preventing Reading Disability: The Reading Readiness Factor," *Educational Administration and Supervision* (1936) 22:401-418.

WRIGHTSTONE, J. WAYNE: "Diagnosing Reading Skills and Abilities in the Elementary Schools," *Educational Method* (1937) 16:248-254.

II. SPELLING

ANDERSON, W. N.: *Determination of Spelling Vocabulary Based upon Written Correspondence*, Studies in Education, Vol. II, University of Iowa, 1921.

ASHBAUGH, E. J.: *The Iowa Spelling Scales*. Bloomington, Ill.: Public School Publishing Company, 1922.

AYRES, L. P.: *Measurement of Ability in Spelling*, Bulletin of the Division of Education. New York: Russell Sage Foundation, 1915.

BREED, F. S.: *How to Teach Spelling*. Dansville, N.Y.: F. A. Owen Publishing Company, 1930.

BROOM, M. E.: *Educational Measurements in the Elementary School*, "Spelling," pp. 95-98, 158-172, "Handwriting," pp. 147-158. New York: McGraw-Hill Book Company, Inc., 1939.

DAVIS, G.: "Remedial Work in Spelling," *Elementary School Journal* (1927) 27:615-626.

ELDRIDGE, R. C.: *Six Thousand Common English Words*. Niagara Falls, N.Y., 1911.

FORAN, THOMAS G.: *The Psychology and Teaching of Spelling*. Washington, D.C.: The Catholic University of America Press, 1934.

GATES, A. I., and RUSSELL, D. H.: *Diagnostic and Remedial Spelling Manual*. New York: Bureau of Publications, Teachers College, Columbia University, 1937.

HILDRETH, GERTRUDE: *Learning the Three R's*. Minneapolis: Educational Publishers, Inc., 1936.

HORN, ERNEST: *A Basic Writing Vocabulary. 10,000 Words Most Commonly Used in Writing*. Monographs in Education, First Series, No. 4. University of Iowa, 1926.

—: "Principles of Method in Teaching Spelling as Derived from Scientific Investigation," *Eighteenth Yearbook of National Society for the Study of Education*. Bloomington, Ill.: Public School Publishing Company, 1919.

—: "Spelling," *Encyclopedia of Educational Research*, pp. 166-183. New York: The Macmillan Company, 1941.

JONES, W. FRANKLIN: *Concrete Investigations of the Material of English Spelling with Conclusions Bearing on the Problems of Teaching Spelling*. Vermillion, S. D.: University of South Dakota, 1913.

PYLE, WILLIAM H.: *The Psychology of the Common Branches*, Chap. VIII. Baltimore: Warwick and York Incorporated, 1930.

THORNDIKE, E. L.: *The Teacher's Word Book*, rev. ed. New York: Bureau of Publications, Teachers College, Columbia University, 1931.

TIDYMAN, W. F.: *Survey of the Writing vocabularies of Public School Children in Connecticut*, Teachers Leaflet No. 15, U.S. Bureau of Education, 1921.

III. HANDWRITING

AYRES, L. P.: *A Scale for Measuring the Quality of Handwriting of Adults*, Russell Sage Foundation Pamphlets. New York: Russell Sage Foundation, 1915.

CONARD, EDITH U.: "Manuscript Writing Standards," *Teachers College Record* (1929) 30:669-80.

FREEMAN, F. N.: "Handwriting," *Encyclopedia of Educational Research*, pp. 555-561. New York: The Macmillan Company, 1941.

— and M. L. DAUGHERTY: *How to Teach Handwriting*. Boston: Houghton Mifflin Company, 1923.

GRAY, C. T.: *A Score Card for the Measurement of Handwriting*, Bulletin No. 37, Austin: University of Texas, 1915.

LEAMER, EMERY W.: *Directions for the Use of the Leamer Diagnostic Practice*

Sentences in Handwriting. Bloomington, Ill.: Public School Publishing Company, 1924.

PRESSEY, S. L., and PRESSEY, L. C.: *The Pressey Chart for Diagnosis of Illegibilities in Handwriting*. Bloomington, Ill.: Public School Publishing Company, 1927.

Teachers Manual, Courtis Practice Tests in Handwriting. Yonkers, N.Y.: World Book Company.

THORNDIKE, E. L.: "Teachers' Estimates of Specimens of Handwriting," *Teachers College Record* (1914) Vol. 15, No. 5.

WEST, PAUL V.: "Remedial and Follow-up Work," *Handwriting (Elements of Diagnosis and Judgment of Handwriting)*, Bulletins No. 1 and No. 2. Bloomington, Ill.: Public School Publishing Company, 1926.

CHAPTER 6

Measurement of Language and Literature

In the educational process, communication through the instrumentality of oral and written language stands at the very top in importance. The struggle to attain effectiveness in the use of language begins when the child speaks his first word and ends only shortly before death. It is a product of his entire milieu and of the inherited capacities which he possesses. Its difficulty increases with age, for the ideas which must be communicated increase in complexity and precision. The simplicity of the sentence structure changes as complex and compound sentences make their appearance. Subtle nuances of thought take the place of gross general statements. Words, too, which at first are largely imitative may themselves become more complicated as usage loads them with meaning. As the child struggles for more correct expression he finds the grammatical forms and words of the home and street different from those of writing and of the school. Sometimes the verbal expressions he hears please him more than the forms that are more grammatically correct. For these reasons the aims and objectives of the teaching of language are difficult to determine.

The following list is a rather incomplete description of these aims but does emphasize the major objectives. These objectives were largely derived from the work of Dora V. Smith.¹

AIMS AND OBJECTIVES OF LANGUAGE TEACHING

1. To teach pupils and students to communicate easily and effectively with others by means of oral and written language.
2. To teach all pupils and students to know and use acceptable forms of language. Written language demands more exact expression and forms more grammatically correct than does oral language.
3. To teach all pupils and students the value of knowing the meaning of many words so as to express more exactly what they have in mind and to understand what others say and write.
4. To teach all pupils and students what a sentence is and some appreciation of the interrelations of words within the sentence. This

¹ Smith, Dora V., "Diagnosis of Difficulties in English," "Educational Diagnosis," *Thirty-fourth Yearbook of the National Society for the Study of Education*, Chap. XIII. Bloomington, Ill.: Public School Publishing Company, 1935.

would prevent the writing of incomplete sentences and would ensure the proper agreement of subject and predicate, the correct use of pronouns, etc.

5. To teach all pupils and students the elements of good taste in writing and speaking through reading and communing with the literature suitable for their age.

6. To teach all pupils and students how to collect materials on a topic, how to organize them, and how to present them effectively.

The special objectives of oral speech are as follows:

1. To attain a special facility with oral forms of speech. If speech habits are almost automatic, opportunity is given for speakers to think while they are speaking. This refers especially to the correct use of pronouns and verbs.

2. To attain the ability to pronounce the words used.

3. To enunciate the words so that they may be understood.

4. To learn how to emphasize particular words and phrases so that certain ideas will stand out.

5. To learn to arrange the gathered material in an orderly manner so that the thought flows smoothly and clearly.

The special objectives of written language are as follows:

1. To learn the mechanics of expression: punctuation, capitalization, and spelling.

2. To develop an understanding of sentence structure and the interrelations of words in a sentence.

3. To master the elements of grammar: agreement of subject and verb, tenses of verbs, pronouns, distinction between words, etc.

4. To learn to become keenly sensitive to words and their usage.

5. To develop a desire to express well and exactly ideas entertained.

6. To attain proficiency in gathering materials, organizing them, and presenting them in writing in an orderly convincing manner. This would include the taking of notes, outlining, and giving attention to the form of presentation.

7. The development of the understanding of the selection, arrangement, and interrelation of sentences within a paragraph so that unity and coherence may be present in the paragraph.

8. To develop a taste for and appreciation of thought beautifully expressed by studying the language of great writers.

9. To attain some proficiency in creative writing.

LANGUAGE TESTS: ELEMENTARY SCHOOL

ORAL LANGUAGE

There are no tests of oral language available at the present time. The author was not even able to find a well-developed check list which

might be applied to oral speech. One not too successful attempt was the use of recording equipment to arrange oral compositions¹ in a scale, but up to the present this procedure has not proved practical.

In E. A. Cross's English Test² there is one section which treats of pronunciation. The directions of this section say:

Place a check mark in the parentheses nearest each correct pronunciation, as in the samples. Give careful attention to the position of the accent mark.

Four examples from this section are:

| | | | | | |
|--------------|----|-------------|-----|-----|-------------|
| 7 recognize | —— | rĕk' à nĭz | () | () | rĕk'ög nĭz |
| 11 salmon | —— | sāl' mŭn | () | () | sām'ŭn |
| 15 regular | —— | rĕg' ū lār | () | () | rĕg' ūl ār |
| 19 sagacious | —— | sā gā' shŭs | () | () | sā gāsh' ūs |

There are 32 words to be pronounced. This test is one of eight in this test and has no separate scores or norms. Its score is added with seven others to make a total for which decile norms are available for grades 8 through 13.

WRITTEN LANGUAGE

Tests for written language cover most of the more formal aspects of language. Considerable attention to language tests has been given by those who construct test batteries. A detailed description of the language tests of three well-known test batteries will be given as illustrations of what most such batteries contain: (1) Metropolitan Achievement Test,³ (2) Stanford Achievement Tests,³ and (3) Iowa Every-pupil Tests of Basic Skills.

The Metropolitan Achievement Tests divides its tests of English into (1) language usage, (2) punctuation and capitalization, and (3) grammar. These language tests are in one sense diagnostic. Before the construction of these tests, careful studies had been made of the most frequent and the most persistent errors that children make in language. For the most part these tests, especially those which test language usage, concentrate on checking these errors. It is assumed that if children know these more difficult aspects of language they will have little difficulty with the rest. In the Metropolitan Achievement Tests, language usage tests appear as early as grades 2 and 3, they continue through grades 4, 5, and 6, and they are developed in a more elaborate form for grades 7 and 8.

¹ Netzer, Royal F., *The Evaluation of a Technique for Measuring Improvement in Oral Composition*, doctoral dissertation, University of Iowa, 1937.

² World Book Company, Yonkers, N.Y., 1923. Items by permission.

³ Items by permission of World Book Company, Yonkers, N.Y.

The form of the language-usage test may be illustrated with a few samples:

Complete the sentence.

1. The baby _____ the milk from the bottle.
2. This is the child w_____ was late.
3. Cats keep clean by washing th_____selves with their tongues.
4. Neither of the two boys w_____ willing to bring it.
5. W_____ do you think will win the prize?
6. "Did he lie down?" "Yes, he l _____ down on the bed and fell asleep."

There are 38 such items at the elementary level and 46 at each of the upper levels. Here are some of the language usages sampled: "give" and "gave," "drank" and "drunk," "may" and "can," "taken" and "took," "sit," "sat," and "set," "lie" and "lay"; past tenses and past participles of such words as "choose," "begin," "grow," "stay," "drive," and "do"; pronouns; "neither . . . nor"; "those kinds"; and many more. Such a large sample of language usage offers many opportunities for the study of the types of difficulty present.

Beginning with grades 4, 5, and 6 there is also a section on punctuation and capitalization. A sample sentence to be punctuated is:

Take it away it is annoying me.

A sample of a sentence in which to enter capital letters and punctuation marks is:

mrs Green is carols aunt.

In the advanced battery (for grades 7 and 8) the more formal aspects of language contain tests of grammar as well as tests of language usage, punctuation, and capitalization. In the test of grammar, tests are arranged for types of sentences, for the number of words in the subject and predicate, and for recognizing several parts of speech. From a short paragraph, also, children are asked to designate simple, complex, and compound sentences. Some attention is also paid to the selection of the one principle among nine which applies to the correct usage of following: "neither . . . nor," "don't" or "doesn't," etc. From this description it is seen that the more formal aspects of punctuation, language usage, and sentence structure are covered. It is also quite clear that *in using the tests for teaching purposes analysis of errors could be studied and special areas of weakness made evident.*

The Stanford Achievement Tests have also provided adequate tests for the formal aspects of language. For example, in the intermediate battery there are 100 items of language usage. Difficulties studied are: "ain't got" and "haven't," "doesn't" and "don't," "did"

and "done," "went" and "gone," "eaten" and "ate," "broke" and "broken," "come" and "came," etc. This test offers a very complete coverage of the usual errors in language usage. In like manner the advanced battery has 100 items of paired words in a sentence, one of which is correct.

I wanted to ^{lay}
lie down and sleep.

Was it ^{her?}
she?

How do you know it was ^{they?}
them?

There is *no attempt* in this set of tests to measure *punctuation, capitalization, recognition of the principles of usage, or grammar.*

In the Iowa Every-pupil Tests of Basic Skills occurs Test C, Basic Language Skills.¹ This test of language skills includes tests of punctuation, capitalization, usage, spelling, and sentence sense. In the test of punctuation there are included sentences with no punctuation whatever. Periods, commas, question marks, quotation marks, and apostrophes are to be properly entered. There is a separate section in which the sentences are properly punctuated but no capitals are used. In the 50 items used for testing language usage, correct and incorrect usage are paired:

There weren't ^{any}
no more nuts.

Kate ^{chose}
choosed the red one.

The test of sentence sense asks the student to place a cross in the R box if it is a good sentence; in the W box if it is not a good sentence.

R W Then as the boys came back to their seats

☐ ☐

R W You are lucky

☐ ☐

There are 40 such sentences. It must be evident that this test covers in considerable detail the more formal aspects of language. Again, there is considerable opportunity for analysis and diagnosis of errors. Most modern test batteries have good sections on language usage, punctuation, and capitalization. In some cases, as in the Metropolitan Achievement Test, there is a thorough coverage of the aspects of language which all pupils should know. Noteworthy also in this respect

¹ Items by permission from Houghton Mifflin Company, Boston.

are the California Achievement Tests and the Coordinated Scales of Attainment. The former of these offers special techniques for analysis of errors.

The author has called attention to these rather well-known matters of analysis and diagnosis because so often testing is regarded as an administrative function. The limitation of the use of tests to administrative functions only is unfortunate. The most valuable uses of testing programs come in the analysis of the results of tests. Test batteries are more frequently given than any other type. In its results resides a gold mine of information about children's learning. If we discover the strong and weak points and arrange materials to overcome these errors, learning will be greatly facilitated.

SEPARATE LANGUAGE TESTS

Tests devoted entirely to assaying language abilities are able to furnish information about many more aspects of language than the test battery. By administering a test battery such as those described in this volume it may become apparent from inspection of the scores that many weaknesses appear in punctuation, spelling and language usage. It may then be decided that a more complete language test is needed.

The Iowa Language Abilities Test is available at two levels: (1) the elementary test, for grades 4 to 7, and (2) the intermediate test, for grades 7 to 10. There are three forms of each test—A, B, and C—to be scored by hand and three forms—Am, Bm, and Cm—to be scored by machines.

The elementary test has five subtests: (1) spelling, (2) word meaning, (3) language usage, (4) capitalization, and (5) punctuation. There are 50 complete items in each subtest and 25 additional items on language usage.

The spelling test consists of recognizing the correct spelling from four spellings, only one of which is correct—the proofreading technique.

10. (1) allready (2) alreddy (3) already (4) already —10 1
2 3 4
46. (1) seperately (2) seprately (3) separatly (4) separately
—46 1 2 3 4

The words selected are those which “seem to present persistent spelling difficulties and are found among words of high social frequency and importance.”¹

The word-meaning subtest (Elementary Test, Form A) offers five

¹ *Manual for Interpreting*, p. 4. Items from this test by permission of World Book Company, Yonkers, N.Y.

choices for selecting the word which (1) means the same as, and (2) means the opposite of the word to be defined.

14. *Fresh* 1 new 2 frozen 3 clear 4 stale 5 cold _____ 14

1 2 3 4 5 1 2 3 4 5
S ■■■■■ O ■■■■■

37. *Counterfeit* 1 genuine 2 new 3 false 4 peculiar 5 contrary
_____ 37

1 2 3 4 5 1 2 3 4 5
S ■■■■■ O ■■■■■

The language usage subtest consists of two parts: (1) correct word forms, and (2) faulty expressions. Correct word forms are tested in the usual way:

12. The hat cost (1) two (2) to dollars _____ 12 1 2

38. The cook (1) ringed (2) rang the dinner bell _____ 38 1 2

Faulty expressions are tested by 25 sentences which are judged "good" or "bad." In the sentences appear quite a variety of sentence errors.

60. The boys they went fishing _____ 60 Good Bad

73. We can easy take two or three in our car _____ 73 Good Bad

The capitalization subtest may be illustrated by the following:

1 2 3 4
7. I hope miss Kelley will come with you _____ 7 1 2 3 4 N

1 2 3 4
23. She felt better, just as dr. Brown said she would _____ 23 1 2 3 4 N

The punctuation subtest checks the correct usage of the period, comma, question mark, apostrophe, quotation marks, and both *s'* and *'s*.

14. The members of the *band* in the meantime, took a recess _____ 14 : , ? N

27. *Dear Edith*

I hope you will come for a visit next week _____ 27 . , s' 's N

The intermediate test of the Iowa Language Abilities Test is constructed much like the elementary test. In both there are spelling, word meaning, language usage, capitalization, and punctuation. The intermediate test adds grammatical form recognition and sentence sense. There are also some minor changes in the manner of constructing the first five subtests. Let us look at the word-meaning subtest. It requires the recognition of a word and its opposite from its given definition:

11. *To have power of endurance*

1 queer 2 sensible 3 selfish 4 strong 5 weak _____ 11

1 2 3 4 5 1 2 3 4 5
S ■■■■■ O ■■■■■

20. *To regard with strong approval*

| | | | | | | | | | | |
|---|--------|---|------|---|---------|---|---------|---|---------|---------|
| 1 | admire | 2 | cars | 3 | delight | 4 | dislike | 5 | advance | _____20 |
| 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 | |
| S | ■ | ■ | ■ | ■ | O | ■ | ■ | ■ | ■ | |

46. *To suffer pain, sorrow, or destructive force patiently*

| | | | | | | | | | | |
|---|--------|---|-------|---|-------|---|--------|---|-----------|---------|
| 1 | exempt | 2 | yield | 3 | annex | 4 | endure | 5 | transport | _____46 |
| 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 | |
| S | ■ | ■ | ■ | ■ | O | ■ | ■ | ■ | ■ | |

The subtest on grammatical form recognition lists eight forms—(1) noun, (2) pronoun, (3) adjective, (4) adverb, (5) verb, (6) conjunction, (7) preposition, and (8) interjection which students are to recognize in 25 sentences. The additional subtest in this intermediate test consists of 50 sentences some of which are complete; others, fragmentary. The test is to mark them "S" if complete; "F" if fragmentary. Examples are: "Birds flying," "The coat hanging in the corner."

Many details of this test have been described to show how completely formal language has been covered. Norms of the test are in grade equivalents and in percentile ranks for grades 4.8, 5.8, 6.8, 7.8, 8.8, 9.8, and 10.8. The care exercised in construction, the number and variety of exercises, the extent of coverage of important details, and the time to administer (48 and 46 minutes) recommend this test for use in the study of language in the elementary school.

Other tests of language usage, capitalization, and punctuation such as the Kirby Grammar Test, Wilson Language Error Tests and Briggs English Form Test have been superseded by the newer tests which have been described in this text.

DIAGNOSTIC TESTS

In spite of the most skillful instruction, errors will creep into the use of language. In the upper grades especially these errors become noticeable to the teacher in the pupil's speech and in his written work. Some of these errors may be detected in the general battery or even more so in the special battery for language abilities. At times, however, a need is felt for a more concentrated analysis and more complete diagnosis of language errors. To meet such a need is the diagnostic test. Satisfactory diagnosis in all areas is impossible, nor has it been attempted. In the area of language usage within the sentence, Franseen's Diagnostic Tests in Language is satisfactory.

Franseen's Diagnostic Tests in Language¹ contains opportunities for diagnosis in three areas: (1) pronouns, (2) verbs, and (3) varied constructions. There are two forms, A and B.

In Part I, 120 different examples of usages of pronouns occur. The

¹ C. A. Gregory Co.

uses of objective or nominative forms in compound subjects, of complements of copulative verbs, of objects of prepositions, of subject of infinitives, of object of a transitive verb are tested. Special errors connected with "us" and "we," with demonstratives, with the use of singular or plural forms in indefinites are included.

Part II, contains 149 opportunities for using correct or incorrect verb forms arranged in the groups. Special attention is given to the use of the past tense or past participle, to the errors arising in the use of present or past tense, to the use of the wrong verbs, and to the use of corruptions, etc.

The third section, varied constructions, emphasizes the agreement of person and number of verbs, errors in the use of adverbs, conjunctions, prepositions, plurals of nouns, adjectives, and ends with 40 items on the recognition of miscellaneous sentence errors.

The tests are scored in terms of errors. No norms or reliabilities are furnished because this test is designed to aid the classroom teacher in discovering errors. Brief suggestions for programs of improvement are furnished for each of the three areas. Each pupil's total score is the number of errors he makes. His achievement may be found by subtracting the sum of his errors from the maximum total score. A class record sheet is furnished with the total score for the different sections of each part printed at the top as well as columns for total pupil error and total pupil achievement.

Two other tests which use the term "diagnostic," perhaps ill-advisedly, are the Leonard Diagnostic Test in Punctuation and Capitalization¹ and the Los Angeles Diagnostic Test in Language.² These tests are too narrowly conceived and too inadequate in sampling to be truly designated as "diagnostic."

LITERATURE

The teaching of literature in the elementary school has as its objective the gaining of acquaintance with writing of good quality in which the ideas are fittingly and sometimes beautifully expressed. It furnishes examples with which the pupil may compare his own expressed experiences and seeks to develop good taste in expression and an appreciation of important ideas well expressed. It aims to develop an interest in good literature and some bases for judging what is and what is not good taste in writing. It is hoped also that the pupil will develop a new sensitiveness to the different use of words in literary expression.

The measurement of literature in the elementary school has up to the present consisted pretty largely of questions about authors and their

¹ World Book Company, Yonkers, N.Y.

² California Test Bureau, Los Angeles, Calif.

works, the identification of a character or quotation with the piece of literature in which it occurred, and the matching of the characters developed in a poem, story, or novel. In general, there has been no attempt to measure powers of discrimination which weigh the characteristics and find this poem or story better than that one. The attempts at measuring have been confined mainly to the general test batteries. The Stanford Achievement Test, Metropolitan Achievement Tests, and the Coordinated Scales of Attainment have well-developed sections on literature. Of these, the Coordinated Scales of Attainment places more emphasis on the facts contained in the story itself than on its author or on the conditions under which it was written. The following items are from grade 6, Coordinated Scales of Attainment:¹

34. Who wandered about doing good, dressed in an old coffee sack, with a stew pan cocked over one ear?
 1 Johnny Appleseed 2 Ichabod 3 Pecos Bill 4 Mike Fink 5 Gareth.
41. In Eugene Fields' poem, "The Duel,"
 1 The characters ate each other up 2 They fought their duel in a garden
 3 They sailed away in a beautiful boat 4 They were stolen away by kidnappers
 5 They made up their quarrel and became good friends again.

This is an item from grade 7:

47. A poem about a soldier who was hanged for shooting a comrade as he slept is
 1 "A Ballad of John Silver" 2 "The Highwayman" 3 "The Skeleton in Armor"
 4 "Danny Deever" 5 "Jim Bludso."

This series of tests emphasizes breadth of reading more than intensity of reading. For example, in grade 7 the test asks what is the story about the Baxter family in Florida, which of five stories is about the life of a musician, and what story (of five listed) pictures life in a family living in Maine.

The section on literature of the Metropolitan Achievement Tests (grades 7 and 8) includes 80 questions about stories and poems. Pupils are asked to finish two well-known quotations, to identify quotations with the poem or story in which they occur, to match characters, to identify leading events with the story, to recognize the country in which the events of the story took place, and to answer many other questions about the contents and characters of stories and poems. Two or three illustrations will make this clear. In the form usually of four multiple choices, children are asked what happened to both arrow and song, for what was the Bell of Atri rung, who was the lawgiver of Israel, in which of four stories was a blue-haired fairy a character, and who was the man who saved Rome. Examples of items are:

¹ Items by permission of Educational Test Bureau, Minneapolis, Minn.

2. Doing nothing is doing _____ 1 too much 2 plenty 3 good 4 ill
() 2
7. At the "Mad Tea Party" the one who was sleeping was the _____ 1 Queen
2 Hatter 3 March Hare 4 Dormouse () 7
17. "The Barefoot Boy" is a poem about _____ 1 an orphan lad 2 a city boy
3 a lazy boy 4 a country boy () 17
38. "The King of the Golden River" tells how _____ 1 a cruel king died 2 a
boy discovered gold in a river 3 a valley became beautiful as a garden 4
Gluck killed a giant () 38

An interesting exercise occurs in Items 56 to 63, which together constitute a matching test. Ten selections are listed, after which appear eight quotations each of which must be identified with the correct selection. For example, "Between the dark and the daylight when the night is beginning to lower," "The breaking waves dashed high on a stern and rock-bound coast," and six other quotations must be matched with the proper selections.

The literature section of the Stanford Achievement Test, advanced battery, contains 50 items almost entirely of the informational variety. Only three short choices are given for each item. When turning to this test from those just considered there is a distinct feeling that this test is more dependent upon memorized details than the others. Three examples are illustrative of this test:

11. Buffalo Bill's name was _____ 4 Kenton 5 Crockett 6 Cody _____
11 4 5 6
:: :: ::
21. The Selfish Giant shut the children out of his _____ 7 house 8 library
9 garden _____ 21 7 8 9
:: :: ::
38. Evangeline lived in _____ 4 Acadia 5 Tuscany 6 Normandy _____
38 4 5 6
:: :: ::

A consideration of the tests of literature present in the three batteries makes it very clear that several aspects of literature instruction are not included. There is no test of the capacity to distinguish between what is good and what is poor in literature. Nor are there specific tests of the organization and development of material within the selection studied. No indication of differential appreciation appears at any point, nor does any evaluation of one's own creative product appear in any test. On the other hand, considerable improvement has been achieved over the older tests by introducing questions concerning the contents of selections rather than by selecting questions as was formerly the custom from the names of authors and their works. Further tests of

LIST OF LANGUAGE AND LITERATURE TESTS—ELEMENTARY SCHOOL

| Name of test | Grades | Kind of test | Number of forms | Types of scores | Publisher |
|---|-----------------|--------------|---------------------|--|---------------------------------------|
| Briggs English Form Test..... | 7-9 | Survey | 2 | Grade scores in percent of errors | Teachers College, Columbia University |
| Franseen Diagnostic Tests in Language | 3-8 | Diagnostic | 2 | Tests scored in terms of errors. Norms regarded as unimportant for diagnosis | C. A. Gregory Co. |
| Hudelson English Composition Scales | 4-12 | Survey | 1 | Grade norms | World Book Company |
| Iowa Language Abilities Tests | 4-6, 7-10 | Survey | A, B, C, Am, Bm, Cm | Standard scores, grade equivalent, percentiles | World Book Company |
| Van Wagenen-Dvorak Diagnostic Examination of Silent Reading Abilities | 4-5, 6-9, 10-12 | Analytical | 3 levels | Raw scores converted to C scores | Educational Test Bureau |
| Leonard Diagnostic Test in Punctuation and Capitalization | 5-12 | Diagnostic | 2 | Grade norms and percentiles | World Book Company |
| Los Angeles Diagnostic Test in Language | 3-9 | Diagnostic | 4 | Grade and age norms | California Test Bureau |
| Cross English Test (high school use here)... | 9-12 | Survey | 3 | Grade norms | World Book Company |
| Pressey Diagnostic Tests in English Composition | 7-12 | Diagnostic | 4 | Grade norms | Public School Publishing Company |
| Nassau County Scale (Trabue)..... | 3-12 | Survey | 1 | Grade norms | Teachers College, Columbia University |
| Van Wagenen English Composition Scales | 3-12 | Survey | 1 | Grade norms | World Book Company |
| Coordinated Analytic Scales of Attainment in Literature..... | 7-8, 9-12 | Survey | 2 levels | Grade norms | Educational Test Bureau |

literature are described when this topic is treated at the high school level.

LANGUAGE TESTS: SECONDARY SCHOOLS

OBJECTIVES AT THE SECONDARY SCHOOL LEVEL

The objectives in teaching language enumerated at the beginning of this chapter are continued in the high school. Mechanics and form, good usage, appreciation, vocabulary development and sentence structure are as important at the high school level as in the elementary school but now they become more complicated. More emphasis is now placed upon the interrelations of sentences within the paragraph and the arrangement of topics within the written composition. The understanding of the structure of language as developed in grammar assumes a position of larger importance. In the area of literature, the discrimination between what is good and what is unacceptable and some understanding of the structure of good writing receive more emphasis.

However, the specific objectives and aims in the teaching of language (English) are often quite difficult to define. Several factors cause this difficulty. The most potent of these, perhaps, is the lack of uniqueness in English work. If all the desirable outcomes of English are listed they coincide almost with the aims of total education. Thus an excellent list of objectives by Dora V. Smith¹ includes such topics as "ability to stick to a subject," "ability to discriminate between important and unimportant material for the purpose in hand," "to give pupils an increasingly adequate vocabulary," and "to develop in pupils the habit of clear, orderly thinking about matters within their own experience." While these objectives are most certainly true of English, they are almost as true of home economics and social science, to select courses at random.

A second difficulty arises out of the subjectivity of some of the most desirable outcomes in the teaching of English. Just what does one *do* differently when he has acquired an appreciation of the poems of Burns or the dramas of Shakespeare? And what English teacher would be satisfied unless his students had an added joy when contemplating some great work of literature?

In the third place, because philosophies differ so widely about what English teaching should be, the very courses themselves contain widely different materials. One teacher, enamored of the past and influenced by hard and fast entrance requirements of some of our great colleges, constructs his reading lists from literary masterpieces that have stood the test of time. Classroom selections chosen to be studied more intensively are of this same type. Another teacher, imbued with a different

¹ Smith, *op. cit.*, pp. 230-233.

philosophy, worships at the shrine of interest. He studies children's interests, tries to understand them and to lead them into new and more realistic areas. Reading lists are made up of those materials that are in demand. Within limits, students make their own reading lists. In the light of such divergent philosophies there is little wonder that materials of instruction vary broadly from course to course. And yet there are some areas where measurement is satisfactory.

When the whole subject of English is considered it is seen that it may very easily be divided into two large areas. One of these areas has to do with *the student's getting acquainted with great literature*. To do this, he must learn how to read literary masterpieces with facility and understanding. This type of reading differs somewhat from ordinary reading because of the nature of the material and its manner of expression. To read literary material with facility and understanding the reader must be highly sensitive to figures of speech, to allegory and symbolism, to classical, mythological, and historical references, etc. Poetry, too, adds other difficulties, for its understanding and appreciation is increased by understanding something of rhyme and rhythm, of poetic license, of the form of the sonnet, of blank verse, and of scansion. When, however, these minutiae have been mastered the full understanding extends one's experiences vicariously, gains for one appreciation and judgment of the best of what has been written, and leads the student to an awareness of our literary heritage—its masterpieces, its authors, its historicity.

The second area in the teaching of English involves *the ability to express one's ideas clearly and effectually* both in speech and writing. Ordinarily, we think of grammar on the one hand and rhetoric and composition on the other. Grammar usually involves a study of sentence structure, spelling, punctuation, and capitalization. At its best, grammar is functional, *i.e.*, its facts are learned in connection with good sentence structure. Composition and rhetoric have more to do with the expression of thought effectively in larger units of the paragraph and the essay or poem. It is thus largely a matter of organization. Oral language and public speaking involve the audience situation. The language used and the manner of expression differ somewhat from those of written language. Unfortunately, no satisfactory measures have been developed in the area of speech. Measurements have been constructed in the following areas:

1. Language structure and usage
 - a. Language usage
 - b. Capitalization and punctuation
 - c. Spelling

- d. Sentence structure
- e. Organization
- f. English composition
- 2. Literature and appreciation
 - a. Reading comprehension and understanding
 - b. Vocabulary
 - c. Literary appreciation, judgment, and acquaintance

LANGUAGE STRUCTURE AND USAGE

Among these categories the tests of language usage have been probably the most satisfactory. The errors in English usage are concentrated for the most part in the agreement of subject and predicate, in the use of irregular verbs, in the correct forms of pronouns, and in the use of a few difficult words.

Tests of English Usage

The grammatical usage division of the Cooperative Mechanics of Expression Test¹ contains 60 sentences with three or four words or phrases in each sentence underlined and numbered. The problem is to recognize the error and place its number in parentheses at the end of the line. Some sentences are entirely correct. Illustrations:

25. He was *in despair*; to *who* could he turn? ()
 1 2 3 4
33. The puppy made itself *at home* and *calmly laid* down near the fire. ()
 1 2 3
37. *Accuracy* of movement, *like* accuracy of words, *are* essential to the *success*
 1 2 3 4
 of magical rites. ()

Differences between "ready" and "already," between "principal" and "principle," are tested. Errors to be corrected in the test involve the use of pronouns as objects of prepositions and verbs as well as correct reference to antecedents.

In the Barret-Ryan-Schrammel English Test, two of the three parts are concerned with language usage. In Part I, Sentence Structure and Diction, correct and incorrect words or phrases in a running account are underlined and the subject must recognize whether the usage is correct or not. Some of the errors present are the use of "of" for "have," lack of parallel construction, use of "affect" for "effect," "set" for "sat," and "most" for "almost." In the second part of this test, Part II, Grammatical Forms, the subject must both recognize the

¹ See lists of tests at the end of the chapter. Items by permission of Educational Testing Service, Princeton, N.J.

error and give its grammatical rule. For example, in "It was left to Jane and I -" the subject must recognize both that "I" is an error and also that it is the object of "to." The subject might even consider a form wrong and give the reason for it, yet be mistaken on both counts. Agreement of verb with subject when phrases come between, proper forms of pronouns, agreement of verb with what comes after "there," recognition of the correct usage of the subjunctive, and the proper use of "myself" are samples of what the test contains. The reliability is indicated by a coefficient of .88 and .89¹ when computed from comparable forms and .91 to .94 when the odds-even technique is used. Furthermore, this test correlates about .74 with final English marks in the first semester of college.

Capitalization and Punctuation

Tests of capitalization usually consist of a sentence or paragraph which requires the subject either to indicate the correctness or incorrectness of the usage set forth or actually to correct the error. In some cases the tests are so constructed that the manuals for the student and for the teacher contain a discussion of the grammar involved. This procedure illustrates beautifully our contention that the major uses of tests inhere in their capacity for diagnoses and analysis of difficulties, followed by the application of the laws of learning at the points of weakness.

Pressey's Diagnostic Tests in English Composition include tests of capitalization and punctuation as well as of grammar and sentence structure. The test of capitalization consists of 28 sentences from which all capitalization has been excluded save that of the first word in the sentence. The subject must write in capitals where they are needed. Since the sample of the usage of capitals was arrived at by means of a study of the frequency with which capitals are used in periodicals, newspapers, and business letters, the discovery of errors in their usage is of the first importance. Two illustrations are:

The Rhine flows from the alps to the baltic sea.

The Children's favorite holidays are christmas and thanksgiving.

After the scoring is done the errors of the children are analyzed. The manual describes the principle of capitalization which has been illustrated in the sentence. For example, in the second of the sentence above, the principle of capitalizing the days of the week, the months of the year, and holidays and church days is illustrated.

This careful analysis of specific errors contrasts rather sharply with the section on capitalization in the Cooperative English Test, Mechanics

¹ *Manual of Directions*, pp. 1-2. Yonkers, N.Y.: World Book Company.

of Expression. This latter test offers a few paragraphs in which correct or incorrect capitalization is to be recognized. Some analysis of errors is possible but there is no provision for the follow-up program available in the Pressey test.

Punctuation, too, varies in the tests from a mere recognition of its being correct or incorrect to the naming of the specific error which is present. In the Barret-Ryan-Schrammel English Test the subject simply recognizes at marked points in a running story whether the punctuation is correct or not. If the underlined punctuation is right he simply marks an R on his answer sheet; if wrong he puts down a W. This test, then, is largely a test of proofreading. In the Cooperative English Test, Test A, Mechanics of Expression, numbers are placed under sections or words where punctuation marks might be placed. The subject then must choose from three alternatives including N (no punctuation) the mark which properly belongs at that number. Unlike the Pressey test the omission need not be properly corrected on the sheet. In general, the most common uses of the comma, semi colon, colon, quotation marks, question marks, and periods are represented. Its reliability is satisfactory and is expressed in terms of standard error of measurement.

Spelling

The objectives of the teaching of spelling in high school are the same as the objectives of the elementary school (page 121). The purpose of teaching spelling is to secure facility in spelling those words ordinarily or most frequently used in written communication. It is also important to establish uncertainty in the spelling of words until the writer is sure that he can spell them correctly. Errors in spelling arise largely because the writer thinks the spelling of a word is correct when it is not and does not feel impelled to look it up. If, however, this uncertainty is too extensive the awareness of possible error in spelling interferes with the easy flow of thought which results in smoothness of composition.

Problems similar to those of the elementary school arise in connection with the administration of the spelling test to high schools. First, shall the words (1) be given embedded in a dictation exercise; (2) be dictated, with their use illustrated in a short sentence, and then dictated again; (3) appear correctly in a variety of misspellings; or (4) appear misspelled among other words which are spelled correctly? All these procedures have been tried out with no final experimental answer as to the greater efficiency of any one method. When a word is dictated, its use is illustrated, and it is then dictated again, the subject's total attention is focused on the spelling process. In ordinary composition attention is divided between the spelling and the ideas being expressed.

The more automatic the spelling procedure, the more attention can be devoted to the thought. Under such conditions the recognition of a misspelled word when the written material is checked looms very large. It would seem then that presenting a misspelled word among others correctly spelled has at least the justification of its use in proofreading.

One of the first serious attempts to construct a spelling test suitable for high schools was Sixteen Spelling Scales.¹ The authors of this scale secured 2,000 most frequently used words from previous studies and from their own experimentation and embodied samples of these into 16 spelling scales of 20 words each. The 2,000 words were submitted in lists of 100 to 46,017 pupils in 181 high schools to be spelled. Each list of 100 was spelled by 160 to 1,200 secondary school pupils. From these data was assembled a list of 2,000 words whose difficulty had been actually determined. From this list 12 lists of 20 words each were arranged in such a manner that the first words in all 12 lists were of equal difficulty, as were the second, the third, etc. Lists XIII through XVI were somewhat more difficult. Each of the 20 words of every test was first pronounced; then the sentence was read to the pupil and the word to be spelled pronounced a second time. Norms (medians) were published for each grade and provision made for the teacher to make his own test by selecting words whose difficulty was known. The strength of such a test depends upon the care with which the words were selected. It is noted that while the word to be spelled was embedded in a sentence it had attention called to it by pronouncing it the second time. The reliability was satisfactory for individual diagnosis provided as many as 100 words were used.

Another test especially prepared for high school students is the Bixler High School Spelling Test, revised edition, for grades 7 to 12.² This is a 63-page booklet which contains 64 40-word lists which may be used for teaching or testing. It contains words from the 5,000 most commonly used words as determined by the Commonwealth investigation. Every word in the test, therefore, is a common word. From this larger list four scales of 100 words each have been prepared for use in high school.

In constructing spelling tests from a larger list the problems of the number of words to be used and their difficulty arise. The number of words to be used depends on the purpose of testing. If the problem is merely that of distinguishing between two grades, then a list of 25 will

¹ Hudelson, Earl, F. L. Stetson, and Ella Woodyard (under the direction of T. H. Briggs and T. L. Kelly), *Sixteen Spelling Scales*, New York: Bureau of Publications, Teachers College, Columbia University, 1921.

² Bixler, Harold E., and Ernest P. Simmons, *Bixler High School Spelling Test*. Atlanta, Ga.: Turner E. Smith Co., 1940.

be adequate. If, however, the question is to measure the spelling capacity of a single child, at least 100 words will be necessary.

Finally, attention may be called to Part III, Spelling, of the Cooperative English Test, Mechanics of Expression. In this spelling test the words were selected from the work of Horn and Ashbaugh. Each word appears misspelled with three other words that are correctly spelled. For example, Item 25 has¹

1. sanctioned
2. receipted
3. registrar
4. parliment
5. none wrong

while Item 26 has¹

1. treatise
2. accessible
3. vengeance
4. embarassing
5. none wrong

It is seen that this is a proofreading type of spelling test.

Spelling scales at the high school level have been less successful than they might have been (1) because the words that all are supposed to know how to spell have not been agreed upon, (2) because the manner of presenting the words to be spelled—whether oral or embedded in writing—has not been certain, and (3) because each person's vocabulary is unique, so that spelling becomes a matter of testing and studying words which the individual himself spells incorrectly.

In closing, it should be recognized that while spelling is truly a part of English it is also an integral part of every other subject in the curriculum and like oral and written English should be the function of the instruction in every area of learning.

Sentence Structure

One of the most important outcomes of English instruction is the ability to write satisfactory sentences. The best indication of this achievement appears in the composition. Next to turning a good sentence is the recognition of one that is well turned. In the Cooperative English Test, Effectiveness of Expression, the subject is given an opportunity to select (1) the most effectively expressed one of two passages, and (2) that one most effectively expressed among four passages. In the case of each judgment of effectiveness he has the additional job of selecting one out of four or five considerations which

¹ Items by permission of Educational Testing Service, Princeton, N.J.

had the most to do with his choice. These choices are furnished at both a lower and a higher level. The following example is from the higher level:¹

Of the four sentences below, which one is most effectively expressed?

1. As the chief was away from home, we were welcomed by his deputy, a ruddy young man with an infectious grin.
2. The chief's deputy was a ruddy young man with an infectious grin who welcomed us because he was away from home.
3. The chief's deputy welcomed us, a ruddy young man with an infectious grin, because he was away from home.
4. The chief was away from home and his deputy welcomed us and he was a ruddy young man with an infectious grin.

Which one of the following considerations had the most to do with your choice of the best sentence in the group above?

1. An adjective clause may be clearer than an appositive.
2. If it is not made clear what a pronoun refers to, the sentence may be ambiguous.
3. Successive clauses connected by "and" may be used when it is desired to give equal weight to various thought elements.
4. Each verb should have a subject.
5. A participle is generally taken to modify the subject of a sentence.

In this test there are five items in which judgments are made between two sentences as to which one is better expressed. There are 10 items resembling in form the illustration just presented. In all cases of choice students must check the consideration which led them to their particular choice. If both levels of this test are used many of the most useful principles of sentence structure are tested. The reliability of the total Cooperative English Test is above .95. In fact, the reliability of each of the parts is in the neighborhood of .95.

Organization

The organization of any written English adds greatly to its effectiveness and clarity of expression. It is one of the outcomes most assiduously sought in rhetoric and composition classes. Attempts have been made to measure this outcome indirectly in one of the divisions, Organization, of the Cooperative English Test, Effectiveness of Expression. One is made for the lower level; the other, for the higher level. Three types of approach have been made in the testing of organization.

The first type sets forth five sentences, one of which does not belong with the other four. In the second type a sentence has been separated into disconnected parts which must be rearranged in the correct order. The following example of the second type is from the lower level:

¹ Items by permission of Educational Testing Service, Princeton, N.J.

- A. The children grow awkward and ruddy
- B. because this is still London
- C. they rush whooping along the cinder tracks
- D. not country children
- E. between the ashbins and straggling flowers
- F. but not sharp-eyed, pallid Londoners either.

From these facts the subject must work out the sequence by saying where *A* would be placed in relation to the others. *B*, *C*, *D*, *E*, and *F* must also be correctly placed. The third type of test presents a partially filled-out outline of a well-known topic and then asks the subject to choose from four options the title which is omitted. The following sample is from the lower level:¹

I (24)

- A. Cleaning the Turkey
 - 1. (25)
 - 2. Removing Pinfeathers
- B. (26)
- C. The Roasting Process
 - 1. (27)
 - 2. Length of Roasting Time

In filling in the incomplete outline above, which one of the following topics would you use for (24) the main heading, I?

- 1. Stuffing
- 2. Preparation for Roasting
- 3. Degree of Heat to Use
- 4. Size of Turkey
- 5. Rinsing Inside of Turkey

In like manner each of the other blanks (25,26,27) has five topics from which to select. The process of organizing materials into an orderly sequence is thus measured by getting subjects to recognize that one sentence does not belong with four others which are grouped around the one idea, that there is a best sequence in separated parts of a sentence, and that a topic has certain recognizable internal relations sometimes called *coherence*.

English Composition

The need of more precision in evaluating English compositions has been felt for a long time. It was thought that possibly a rating scale might achieve at least some of the precision desired. After Thorndike had demonstrated in 1909 that the Cattell-Fullerton theorem of equally often noticed differences could be applied to *general merit* in a hand-

¹ By permission of Educational Testing Service, Princeton, N.J.

writing scale, Milo B. Hillegas applied the same principle to constructing a Scale for the Measurement of Quality in English Composition for Young People which was published in 1912. This scale was composed of samples of compositions varying by known units from very poor to very good. The known units were about one probable error apart, a fact derived from the consideration that 75 per cent of competent judges chose one sample as being better than another. The users of the scale simply slid a child's composition along the scale until its general merit equaled that of a sample on the scale; its score, then, was that of the sample. This scale has been improved in certain particulars. In the first place, the compositions which composed the scale were on different topics, which made comparison difficult. Trabue corrected this weakness by building the Nassau County Scale on the same general principle but requiring that all the samples must be written on the topic "What I Should Like to Do Next Saturday." It was soon recognized that children wrote better compositions when they wrote of familiar experiences or on topics about which they were well informed. Lack of information on a topic, therefore, produced poorer quality in compositions. Hudelson's composition scale attempted to overcome this difficulty by furnishing the data from which the composition was to be constructed. The children simply had to retell Aldrich's "A Snowball Fight on Slatter's Hill" after it had been read to them. Some experimenters thought also that if this "general merit" were broken down into smaller parts and those parts combined, more precise measurement could take place. As a consequence, Van Wagenen constructed a set of scales composed of scales for (1) exposition, (2) narration, and (3) description. Each composition was to be rated three times: (1) once on thought content (*t*); (2) once on structure (*s*); and (3) once on mechanics (*m*). The combination was made by using this formula:

$$GM \text{ (general merit)} = \frac{4t + 2s + 1m}{7}$$

This procedure looked efficient but did not work out so well in practice because the errors of rating were probably additive. At any rate, the reliability of the scale's application was no higher than when only general merit was rated. Lewis narrowed the field greatly by constructing a scale made up of letters used in daily life—mail orders, letters of application, social letters, etc. It emphasized the same principle of construction as the Hillegas scale.

Good usage of the scales demands that the student go through a rather rigid training in their use. After such practice more reliable results can be obtained in judging compositions. Generally speaking,

the fundamental difficulty lies in the process of rating itself. For the reliable rating of any composition at least three raters—preferably, five or six—would be needed for accurate results. For these reasons, composition scales are very little used at the present time. The writer believes, however, that for survey purposes in indicating the general level of composition ability of a class such a scale as the Hudelson's could be of real service. This scale would indicate the real level of achievement more nearly than the present rating schemes because the usual level for that grade would be before the rater. For example, quality 4.7 was found by Hudelson's to represent the median performance of children in grade 7. Medians for other grades are 3.6 for grade 5 and 4.2 for grade 6.

Here is the composition which most nearly represents seventh-grade achievement.¹ It is a trifle better than the average for grade 7.

A Snowball Fight on Slatter's Hill

It was on Slatter's Hill that the Battle took place. Slatter's Hill is the boundary line between the North End and the South End.

We took possession of the hill one afternoon and made us a fort of snow. Under the command of Colonel J. Harris we made plenty of ammunition. Some three hundred snow-balls.

The South End was enraged when they saw what had happened and the silk handkerchief that floated on the flagstaff waved defiance to the enemy. They resolved to attack the fort that afternoon and under the brave and daring command of Mat Ames they climbed the height. They were slowly advancing toward our strong hold while we lay in wait.

Each man was well supplied and the orders were not to be sparing with the ammunition. As Ames led his men nearer and within range of the fort. Our noble commander jumped upon the breastworks and took daedly aim at the advancing officer of the enemy.

The aim was fatal for the spinning snowball hit its aim and the enemy's leader went rolling down the hill.

This confused the enemy and our captain took advantage of the situation and ordered rapid firing on them. This being done the enemy was soon put to flight except a few who were climbing the breast works. And they were captured.

One of the strong points of the Hudelson scale is the set of prejudged exercises which the user can practice on. One can thus note each of these samples and compare his rating with the established

¹ By permission of Public School Publishing Company, Bloomington, Ill.

scale values. Constant errors of overrating or underrating can then be corrected.

The composition ability scale does offer levels of attainment usually reached by the averages of the various grades. It thus furnishes an attainable goal toward which pupils may strive. When the compositions seem absolutely hopeless to the college-trained teacher of English he can look at the sample for his grade and be comforted. This goal, since it is within reach, may become a stronger motivating influence than one which approaches perfection.

READING COMPREHENSION AND UNDERSTANDING

Another factor which is related to English as well as to other subjects is reading. For many years instruction in reading was left almost entirely to the elementary school. But when analyses of failures in both high school and college were made it was discovered that poor reading was frequently the cause. Reading failure in most cases revolved around the problem of comprehension. Students could not read the texts which were assigned to them either because these books were too heavily laden with unknown words or because they had never been able to put the parts of sentences together in their minds into a meaningful whole. For these and other reasons there is today in most good high schools a definite objective aimed at improving comprehension in reading.

Many tests have been constructed to test reading for understanding. Some of them have been content to ask questions which might be answered by copying the correct answer verbatim from the paragraph. Others, and these are the best, have included many questions which could be answered from an understanding of the paragraph as a whole. Difficulty has been controlled by increasing the subtlety of the questions and by increasing the complexity and vocabulary of the paragraph.

One of these tests with forms suitable for both the junior high school (lower level) and the senior high school (upper level) is the Cooperative English Test, reading comprehension. The first part of this test consists of 60 words to be defined, but the second, which requires 25 minutes to take, is a test of reading. In the test suitable for the junior high school, 19 short literary selections of somewhat increasing difficulty constitute the material to be read. Four or five questions are asked about each paragraph. An illustration makes clear the technique:¹

September 3rd (Lord's Day)—Up; and put on my colored suit very fine, and my new periwig, bought a good while since, but durst not wear; because the plague was in Westminster when I bought it; and it is a wonder what will be the fashion after the plague is done, as to periwigs, for nobody will dare to buy any hair, for fear of the infection, that it had been cut off the heads of people dead of the plague. To

¹ By permission of Educational Testing Service, Princeton, N.J.

church, where a sorry dull parson, and so home and most excellent company with Mr. Hill and discourse on music.

82. This passage is apparently taken from

1. an essay
2. a diary
3. a novel
4. a short story
5. a sketch

82()

83. The writer had been afraid to wear the new periwig because

1. he did not know whether it was still fashionable
2. he feared that it was improper in time of plague
3. wigs are easily infected
4. it had come from the hair of plague-stricken persons
5. it was bought in a plague-stricken area

83()

84. The writer may best be described as a

1. tenderhearted person
2. timid person
3. music lover
4. practical person
5. scoffer at religion

84()

85. The tone of this passage is

1. ironical
2. persuasive
3. solemn
4. emotional
5. matter of fact

85()

It will be noted that while this test includes literary selections there is no poetry to be comprehended. This test has high reliability and covers the testing well when only short paragraphs are to be read. Such tests should have at least one or two passages long enough to test the understanding arising out of the interrelation of paragraphs.

But the comprehension of reading literature involves more than a simple understanding of what is stated. What is written may be satirical or merely fanciful. The whole passage may have as its principal purpose the creation of a mood in the reader such as that created in "The Fall of the House of Usher." The understanding of figures of speech, poetic license, references to Greek mythology, the verse form, rhythm and of much else is involved in comprehending a literary selection. In brief, reading literature has certain peculiarities of its own. For this reason, we have such tests as the Cooperative Literary Comprehension Test which uses as its reading material selections from prose and poetry of high literary value. One of the questions usually asked about these selections is the mood conveyed. Fourteen selections, varying from six

lines to about a half a page in length, form the materials for reading for understanding. Here is a short selection with its questions:¹

The sky is low, the clouds are mean,
A travelling flake of snow
Across a barn or through a rut
Debates if it will go.

A narrow wind complains all day
How someone treated him;
Nature, like us, is sometimes caught
Without her diadem.

The central thought of the poem is that

- 10-1 Nature and people have more than one aspect
- 10-2 Winter is depressing
- 10-3 Winter comes upon us suddenly
- 10-4 The wind is very tiresome 10()

The day described is

- 11-1 invigorating
- 11-2 depressing
- 11-3 frightening
- 11-4 soothing 11()

In sound the wind is

- 12-1 howling
- 12-2 hustling
- 12-3 murmuring
- 12-4 whining 12()

The last two lines suggest that

- 13-1 nature does not always seem sublime
- 13-2 nature is sometimes caught unawares
- 13-3 nature does not always rule supreme
- 13-4 the night is occasionally starless 13()

Two scores may be obtained: (1) speed-of-comprehension score, and (2) level-of-comprehension score. The first of these "represents the product of the rate at which an individual has attempted to compre-

¹ By permission of Educational Testing Service, Princeton, N.J.

hend the test material and his success in comprehending it." The second score "provides a measure of the ability of the student to understand the meaning of poetry and literary prose and of his familiarity with literary devices and modes of expression."¹ Percentile norms are available both at the high school and college level. The accuracy of measurement or reliability uses the standard error of measurement. As for all other Cooperative tests, comparisons are made on the basis of scaled scores. There are three forms of the test. The reliability is very high. For Form 0 the reported coefficient is .97.

Probably the most used test of high school reading is the Iowa Silent Reading Tests, advanced tests. It has already been described in this text (page 111). It is mentioned here because the passages to be read are comparatively long and include poetry as well as science and government. It also has good questions on selecting the topic of a paragraph. Its vocabulary test and its test of abilities to look up facts in an index measure useful reading functions. From its part scores, analysis of reading capacities may be made.

The following are a few reading tests suitable for high school students. The one to be used depends much on the problem being attacked.

1. Nelson Denny Reading Test. Houghton Mifflin Company, Boston.

2. Pressey Reading Tests. Ohio State Department of Education, Columbus, Ohio.

3. California Reading Tests, grades 7-13. Intermediate, grades 7-9; advanced, grades 9-13. California Test Bureau, Los Angeles, Calif.

4. Traxler Reading Tests. Public

School Publishing Company, Bloomington, Ill.

5. Van Wagenen Reading Scales. Educational Test Bureau, Minneapolis, Minn.

6. Van Wagenen-Dvorak Diagnostic Examination of Silent Reading Abilities, grades 6-12. Junior division, grades 6-9; senior division, grades 10-12. Educational Test Bureau, Minneapolis, Minn.

Of this list the most diagnostic is the one by Van Wagenen.

VOCABULARY TESTS

Vocabulary tests form a part of many of the tests of English grammar and usage. The knowledge of words and their meanings is also directly related to the measurement of the growth of intelligence. In the original Stanford-Binet, in the Terman-Merrill Revision, in the Wechsler-Bellevue, and in the vast majority of the verbal group tests, vocabulary tests have been found to furnish useful items for measuring intelligence. The present discussion is an attempt to illustrate and describe some of the vocabulary tests which are useful in their own right.

The major problem in constructing vocabulary tests is the selection

¹ *Manual*.

of representative words. In the Cooperative Vocabulary Test¹ there is a sampling from many subject-matter fields. All were selected from Thorndike's *Readers Word Book of Twenty Thousand Words*. Thirty-six of the words were less frequently used than the 20,000. Altogether there are 210 words to be defined. Percentile norms are furnished for public secondary schools of the East, Middle West, and West (school systems of 12 grades) and public secondary schools of the South (11 grades at that time). The test may be given without a time limit. The test's reliability is satisfactory. Here are two samples at the more difficult level (Form V):

22. *candor*

- 22-1 charm
- 22-2 personality
- 22-3 tact
- 22-4 frankness
- 22-5 logic

27. *chimerical*

- 27-1 fantastic
- 27-2 doubtful
- 27-3 temporary
- 27-4 bell-like
- 27-5 synthetic

Another vocabulary test is the Inglis Tests of English Vocabulary.² The words of this test represent a truly random sample of the intelligent general reader's vocabulary. Experiments showed that 150 words were necessary to secure a reliable test. If more than 150 words were used the reliability was not greatly increased. There are three forms—A, B, and C—each of which has a reliability in the neighborhood of .90. Two illustrations are:

| | | | | | |
|--------------------------------|-------------|--------------|-------------|---------------------|---------------|
| He <i>propitiated</i> them | (1) evicted | (2) assisted | (3) praised | (4) ap- peased | (5) angered |
| He <i>uttered</i> the document | (1) wrote | (2) read | (3) recited | (4) dis- covered | (5) published |

Other tests of word knowledge suitable for the high school are:

1. English Vocabulary Tests for High School and College Students. Author: W. T. Markham. Public School Publishing Company.
- The Thorndike Test of Word Knowledge. Teachers College, Columbia University, New York.

¹ Cooperative Test Division, Educational Testing Service, Princeton, N.J. Items by permission.

² Ginn & Company, Boston. Items by permission.

LITERATURE AND ITS APPRECIATION

LITERARY JUDGMENT, ACQUAINTANCE, AND APPRECIATION

One of the most important outcomes of instruction in the teaching of English and the most difficult to measure is literary discrimination. This quality involves two processes: one of them is the capacity to distinguish between what is good in literature and what is merely sentimental, cheap, or tawdry; the other is an acquaintance with and a knowledge of what is generally described as good literature. The latter aspect has been more accurately measured than the former. The measurement of discrimination has been attempted by judging which one is the best of four samples and which one the worst. Difficulties arise here because the best sample of poetry is usually selected from poems generally regarded as good literature. In this case, the subject may choose as best that sample which he has once studied. Some subjects, then, do not make a judgment but simply agree with the judgment of others. Scores on such a test therefore are a mixture of true discrimination and memory. This phase of English instruction has not been well measured up to the present.

MEASUREMENT OF APPRECIATION OF LITERATURE

The meaning of the process of appreciation involves both a feeling tone and a judgment of value. This affective coloring which is added to the judgment is aroused by excellencies in both form and thought. Appreciation consists of "Emotional responses which arise from basic recognitions, enhanced by an apprehension of the means by which they are aroused."¹ Pooley believes that we should identify clearly the objectives of appreciation, prepare items which test them, and then validate the items. He divides appreciation of poetry and prose into two parts: fundamental and secondary. The fundamental responses in poetry arise out of recognition of rhythm, of meter, the grouping of sounds, and the relation of sound to sense (onomatopoeia). Secondary responses in poetry come largely from the comprehension of the content. Among these responses, continues our author, are those arising out of the recognition of emotional overtones, poetic diction, figures of speech, literary allusions, and literary patterns such as verse forms, blank verse, and sonnets, and finally the response arising out of personal experience. In prose, fundamental responses of appreciation arise from the perception of variety in sentence structure and word order and in the length of sentences. The effect of the sequence of sounds and of the

¹ Pooley, Robert, "Measuring the Appreciation of Literature," *English Journal* (High School Edition) (1935) 24:627-633.

grouping of sounds is in the order of appreciation. The recognition of orderly time and space progression adds something to the fundamental appreciation. Secondary appreciative responses come to the individual when he is aware of the means by which all the fundamental responses are aroused. More specifically, appreciation is enhanced when the subject recognizes the relationship between word order, sentence structure and the content of the material, the appropriateness of the choice of words to the content, and the figures of speech and when he identifies himself with the characters portrayed.

Other aspects of appreciation have been emphasized by the staff of the Progressive Education Association. Here are the headings of the overt acts and verbal responses which are illustrated with appropriate subheads in the text.¹

1. Satisfaction in the thing appreciated.
2. Desire for more of the thing appreciated.
3. Desire to know more about the thing appreciated.
4. Desire to express one's self creatively.
5. Identification of one's self with the thing appreciated.
6. Desire to clarify one's own thinking with regard to the life problems raised by the thing appreciated.
7. Desire to evaluate the thing appreciated.

There has been no measure of appreciation developed which attempted to analyze out and then test the elements of which it is composed. Most attempts at measurement have been content to offer an opportunity to choose from selected poems or prose selections the best one and the worst one or to rank them in order. Samples of these attempts are now presented for both prose and poetry.

One of the most interesting attempts to measure the ability to judge the quality of poetry *Exercises in Judging Poetry*, was developed by Abbott and Trabue in 1921. A poem written by a recognized poet was rewritten with varying degrees of literary merit. The instructions were: "Read the poems A, B, C, D, trying to think how they would sound if read aloud. Write 'Best' on the dotted line above the one you like best as poetry. Write 'Worst' above the one you like least." Here is an example:²

Set 13. The Fog

A (.....)

The fog comes
on little cat feet.

¹ Smith, Eugene R., Ralph W. Tyler, *et al.*, *Appraising and Recording Student Progress*, pp. 251-252. New York: Harper & Brothers, 1942. By permission.

² By permission, Bureau of Publications, Teachers College, Columbia University, New York.

It sits looking
over harbor and city
on silent haunches
and then moves on.

B (.....)

The fog is as
quiet as a cat.
It comes creeping over
the city
and stays there quietly until the
first thing you
know it is gone.

C (.....)

The fog is like a maltese cat,
it is so gray and still,
and like a cat it creeps
about the city streets.
How gray it is! How cat-like!
Especially when it steals away,
Just like a cat.

D (.....)

Who sends the fog
so still and gray?
I fondly ask.
And Echo answers,
"E'en the same all-seeing Eye
that sends the still, gray cat."

There are altogether 13 groups of four, of which the sample above is the most difficult.

In M. G. Rigg's, *Measuring the Ability to Judge Poetry*, comparison is made between two samples in each item. Forty pairs of samples are to be judged. The instructions say, "Below you will find some selections of poetry arranged in pairs. For each pair, place an X before the selection which you regard as the better poetry." In each pair of samples one is taken from the works of a recognized poet; the other is a parody of it. The correct scoring was done by 47 college professors, 43 of whom were professors of English. Two samples follow:¹

¹ By permission of Bureau of Educational Research and Service, University of Iowa, 1942.

- 10 A(_____) The night was still. You could not hear the howls
 Of any birds or any bats or owls.
 B(_____) You could not hear, I thought, the voice of any bird,
 The shadowy cries of bats in dim twilight
 Or cool voices of owls crying by night.
- 31 A(_____) Who shall declare the joy of the running!
 Who shall tell of the pleasures of flight!
 B(_____) Oh what a joy there is in running!
 And what pleasures there are in flight!

The reliability coefficient of Form C with Form D is .72.

Other tests of appreciation suitable for high school are:

- | | |
|---|--|
| <p>1. Logasa and Wright Tests for Appreciation of Literature. Public School Publishing Company, Bloomington, Ill.</p> | <p>Tests. Turner E. Smith and Company, Atlanta, Ga.</p> |
| <p>Cook-Bixler Literary Appreciation</p> | <p>Cooperative Literary Comprehension and Appreciation Test. Cooperative Test Service, New York.</p> |

The outstanding attempt to measure appreciation in the realm of prose is the Prose Appreciation Test by Herbert A. Carroll. Tests are now available for (1) the junior high school, (2) the senior high school, and (3) college. Validity is claimed for this test on two counts: the manner of selecting the material and the procedure used in validating the selections. It was at first decided that four selections of differing degrees of literary merit were to constitute each item. Each selection was to be about 100 words in length. All first choices were selected from authors of the highest ability (Tolstoi, Cather, Conrad). The second choices were selected from writers considered second-class (Harold Bell Wright, Ernest Poole, Temple Bailey). Third choices came from magazines with little or no literary merit (*Wild West Weekly*, *Love Story Magazine*, etc.). Fourth choices were deliberate mutilations of the first choice. These choices were further validated by submitting them to be voted upon by (1) members of university English staffs, (2) critics and authors, and (3) high school teachers of English. Only items agreed upon by this galaxy of judges were retained.

A test item consists of four prose selections of about 100 words each, differing rather radically in the amount of literary merit which each contains. There are 10 sets with four selections each at the junior high school level, 12 at the senior high school level, and 14 sets at the college level.

The instructions and the space for rating the selections appear as follows (college level):¹

¹ By permission of Educational Test Bureau, Minneapolis, Minn.

Here is an illustration of the selections to be rated:¹

A MAN

A

He had come to Africa, one might have said, without a face—with only a soft, embryonic boyish countenance upon which life had left no mark; but now, at twenty-six, his features were hardened and sharpened—the straight, rather snub nose, the firm but sensual mouth, the blue eyes in which a flame seemed to be forever burning. The fevers left their mark. There were times when, dead with exhaustion, he had the look of a man of forty. Behind the burning eyes there was forming slowly a restless, inquiring intelligence, blended oddly of a heritage from the shrewd woman who was always right and of the lanky cleverness of a father he could not remember.

B

Dion Taylor was less than thirty. But he was a hundred years older than Cecilia in soul. He was handsome, brown-haired, tall, "taller than Pop and fully as tall as Tom," Cecy had already decided. He had laughing brown eyes and a sophisticated mouth. He wore his evening clothes as nobody else in the room could wear them and his conversation smacked of the world: colleges, ocean liners, studios in Paris and New York. He was rich, he associated only with nice people, and was the youngest member of a very young firm of brokers.

C

Peter was as handsome a fellow as a girl would hope to meet. He was tall and broad shouldered, with eyes as blue as summer skies, hair black as a raven's wing, lips as red as red roses, teeth white as milk, skin brown as a nut, wonderful hands, long legs, a wonderful nose, the best-looking build, walked like a soldier, and had a wonderful voice. He was a prince among men, that was what Peter was. He was so handsome he ought to have been in the moving pictures. Everybody knew this.

D

He was only thirty and he was tall and as fair as Diana was dark; he was amusingly sophisticated and so rich that he never had to think about money. He bought his clothes in London, his wines in France, his automobiles in Italy . . . His gray tweeds greatly became him. His eyes were blue and just the shade of blue she most admired. His hands were nice and brown and well shaped. His voice was the correct sort of voice and he smelt of good tobacco and a certain brand of eau de cologne.

¹ By permission of Educational Test Bureau, Minneapolis, Minn.

Percentile norms based on 200 to 500 cases have been prepared for each of three levels. Probably the greatest weakness of the test is its reliability which, whether computed by the split-half method or that of test-retest, turns out to be .71.

LITERARY ACQUAINTANCE

Literary acquaintance offers, too, its difficulties when measurement is undertaken. The simpler more superficial facts such as authorship of poems, leading characters in a play or novel, or the identification of some striking incident or quotation lend themselves rather easily to objective testing. On the other hand, the details of character development, the unfolding of a complicated plot, the aesthetic appreciation obtained from the literary selection as a whole have thus far escaped measurement. English teachers thus object at times to tests of acquaintance for fear that these tests will influence the teaching of the facts about a masterpiece rather than its inner essence.

The Cooperative Literary Acquaintance Test has three parts:

| Part | Items |
|--|-------|
| I. Pre-Renaissance and Foreign..... | 30 |
| II. English and American, 1500-1900..... | 90 |
| III. Modern English and American..... | 30 |

In Part I the subject is asked to identify Echo, Job, Loki, Charm, Pegasus, Grendel, Terpsichore, and Utopia. He is asked the name of Robin Hood's sweetheart, the Biblical character who sacrificed his birthright, and who of five named authors influenced the drama most. In Part II similar questions are asked about how Lochinvar won his bride, what Pepys is known for, where Pandemonium was located and the location of *Old Creole Days*. In Part III such questions are proposed as the name of the city around which action revolved in *Gone with the Wind*, what *North to the Orient* is about, who Emperor Jones was in the play of that name, and what *Tobacco Road* deals with. Teachers of literature challenge the sampling of the world of literature which this or any other test makes. The questions in all three parts are couched in the form of the following sample.

27. The poet who most interested Amy Lowell was

1. Shelley
2. Swinburne
3. Arnold
4. Keats
5. Byron

| Name of test | Grades | Kind of test | Number of forms | Kinds of scores | Publisher |
|--|-------------|-----------------------|-----------------|--|--|
| Barrett-Ryan-Schrammel English Test. | 9-12 | Survey | 3 | Percentile ranks | World Book Company |
| Bixler High School Spelling Test, Revised. | 9-12 | Survey | 4 | Words spelled correctly Scaled scores | Turner E. Smith & Co. |
| Cooperative English Test | | | | | |
| A. Mechanics of Expression | | | | A. Total score only | |
| B. Effectiveness of Expression | | | | B. Total score only | |
| C. Reading Comprehension. | 7-12 | Survey | 4 | C. Four separate scores | Cooperative Test Service |
| Sixteen Spelling Scales (Huddelson, Stetson, Woodyard) | 7-12 | Survey | 16 | Grade norms | Teachers College, Columbia University |
| Hillgas Scale for the Measurement of Quality in English Composition for Young People | 4-12 | Survey | 1 | Grade norms | Teachers College, Columbia University |
| The Cooperative English Test—C ₂ , Reading Comprehension. | 11-12 | Survey | 4 | Scaled scores | Cooperative Test Service |
| The Cooperative Literary Comprehension. | 10-12 | Survey | 3 | Scaled scores | Cooperative Test Service |
| Nelson-Denny Reading Test. | 9-16 | Survey | 2 | Grade norms | Houghton Mifflin Company |
| Pressey Reading Tests. | 12 | Survey | 1 | Grade norms | Ohio State Department of Education |
| California Reading Tests. | 7-9, 9-13 | Survey (and analysis) | 3 | Grade norms and percentiles | California Test Bureau |
| Traxler Reading Tests. | 7-10, 10-12 | Survey | 4 | Percentile norms | Public School Publishing Co. |
| Cooperative Vocabulary Test. | 7-12 | Survey | 2 | Percentile norms | Cooperative Test Service |
| Michigan Vocabulary Profile Test. | 9-12 | Survey | 2 | Grade norms | World Book Company |
| Inglis Tests of English Vocabulary. | 9-12 | Analytical Survey | 3 | Grade norms | Ginn & Company |
| Abbott and Trabue Exercises in Judging Poetry. | 9-12 | Survey | 2 | Not well standardized | Teachers College, Columbia University |
| Rigg Measuring the Ability to Judge Poetry. | 2 | Survey | 3 | Number of items right | Bureau of Education Research and Service, University of Iowa |

The norms given on basis of entering freshmen, sophomore, junior, and senior are all at the college level.

Another difficulty arises in selecting tests of literature for a particular school because the test items may be unfamiliar to these children. Tastes of English teachers are so different that the selections studied are widely varied. It is thus of great practical importance for the teacher to have a hand in the selection of the test to be used, for, above all, a test must have curricular validity.

Other tests of literary acquaintance are:

- | | |
|---|---|
| 1. Smith and Bixler. Awareness Test of Twentieth Century Literature. Turner E. Smith Co., Atlanta, Ga | Kansas State Teachers College, Emporia, Kans. |
| 2. Barrett-Ryan Literature Test. Bureau of Educational Measurements, | 3. Analytical Scales of Attainment in Literature, grades 7-8, 9-12. Educational Test Bureau, Minneapolis, Minn. |

SUMMARY

The measurement of objectives of language teaching has been most successful in the areas of written language. For the elementary school there are tests of language usage, punctuation, capitalization, and spelling. The study of the most common and most persistent errors of grammar has made it possible to include sentences in which the correct and incorrect forms appear together. Tests for recognizing good or bad sentences, the proper use of pronouns and verbs, and parts of speech are available.

In the high school, tests of the more formal aspects of language such as those of punctuation, spelling, capitalization, and language usage are continued at a higher level of complexity. There are also tests for the proper arrangement of sentences in a paragraph, and scales for rating English composition. Tests for the more formal aspects of language instruction are successful and highly useful.

The measurement of literary appreciation, literary discrimination and sensitiveness to literary expressions is less satisfactory. These qualities are supposed to be developed through the study of literature. The greatest difficulty in measuring literary understanding and appreciation is to avoid superficial aspects of authorship and literary acquaintance and to test for the true meaning and significance present in a poem or story. An inspection of the tests suitable for the elementary school will convince one that this difficulty has been only partially met. In the elementary school these tests of literature are made up of various types of matchings: (1) of authors and their works, (2) of two characters that appear in the same selection, and (3) of a quotation and the poem or book where it occurred. Some items ask about the content of a poem or story; others ask that well-known quotations be com-

pleted; while still others ask that a character be recognized from his description. Elements of good taste in writing are measured by scores on the test of language usage. There are few or no tests of the ability to discriminate between good and poor poetry, nor are there any tests of the organization and development of thought except for scales of composition in the upper grades.

In the high school, provision has been made for testing the capacity to read and understand literary selections. Tests of literary discrimination, both of poetry and prose, have been constructed, although these tests are not sufficiently well standardized to warrant much confidence in them. Tests for the organization and development of the paragraph by recognizing the proper sequence of sentences seem promising.

QUESTIONS AND EXERCISES

1. *a.* Describe the objectives of language instruction.

b. Which general test battery seems to you to measure language usage best?

c. Plan out a testing program as a preliminary procedure for an attack upon a general program of language improvement.

d. Why are objectives in the teaching of English so difficult to define? How does theory enter into your explanation?

2. *a.* Describe the subjective outcomes of English instruction.

b. Have they been well measured by standard tests? Explain.

c. Describe the tests used in testing literature in the elementary school.

3. Evaluate the tests of English usage at the high school level. Why are objectives relating to usage more easily measured than those of appreciation or judgment?

4. What procedures are used in measuring capitalization, spelling, and

punctuation? Which ones are to be preferred?

5. How does the measurement of ordinary reading differ from that of measuring literary selections including poetry?

6. What difficulties appear in the measurement of organization? In your judgment how effective is the test of organization?

7. On what principle was the first composition scale constructed?

8. What factors are to be overcome in measuring the understanding of literary selections which do not appear in the understanding of selections not of a literary nature?

9. Compare the vocabulary tests mentioned as to (*a*) manner of selecting the words, (*b*) the arrangement of words, and (*c*) the reliability of the different tests.

10. What effect might a test of literary acquaintance have on the teaching of literature? How can this danger be avoided?

BIBLIOGRAPHY

Books

GREENE, HARRY A., ALBERT N. JORGENSEN, and J. RAYMOND GERBERICH: *Measurement and Evaluation in the Secondary School*, Chap. XIV. New

York: Longmans, Green & Co., Inc., 1943.

HAWKES, HERBERT E., E. F. LINDQUIST, and C. L. MANN: *The Construction and Use of Achievement Examina-*

tions, Chap. VIII. Boston: Houghton Mifflin Company, 1936.

ODELL, C. W.: *Education Measurement in High School*, Chaps. IV, V. New York: Appleton-Century-Crofts, Inc., 1930.

REMMERS, H. H., and N. L. GAGE: *Educational Measurement and Evaluation*, pp. 33, 214, 302-304. New York: Harper & Brothers, 1943.

ROSS, C. C.: *Measurement in Today's Schools*, pp. 46-49. New York: Prentice-Hall, Inc., 1947.

SMITH, EUGENE R., RALPH W. TYLER, et al.: *Appraising and Recording Student Progress*, pp. 246-276. New York: Harper & Brothers, 1942.

SYMONDS, P. M.: *Measurement in Secondary Education*, Chap. V. New York: The Macmillan Company, 1927.

TRAXLER, ARTHUR E.: *Techniques of Guidance*, pp. 78-81. New York: Harper & Brothers, 1945.

Articles

CARROLL, HERBERT A.: "A Method

of Measuring Prose Appreciation," *English Journal* (1933) 22: 184-185.

———: "A Standardized Test of Prose Appreciation for Senior High School Pupils," *Journal of Educational Psychology* (1932) 23: 401-410.

LOGASA, HANNAH, and MARTHA JANE MCCOY: "Tests for Measuring Appreciation," *School Review* (1925) 33: 491-492.

POOLEY, ROBERT: "Measuring the Appreciation of Literature," *English Journal* (High School Edition) (1935) 24: 627-633.

RIGG, MELVIN G.: "Measuring the Ability to Judge Poetry," *Proceedings of the Oklahoma Academy of Science* (1939) 19: 157-158.

SMITH, DORA V.: "Diagnosis of Difficulties in English," "Educational Diagnosis," *Thirty-fourth Yearbook of the National Society for the Study of Education*, Chap. XIII, pp. 220-233. Bloomington, Ill., Public School Publishing Company, 1925.

CHAPTER 7

Measurement of the Social Sciences

Measurement in the social sciences has been retarded because of a failure on the part of curriculum makers to agree upon desired end results of social-studies teaching, and because of the difficulty of measuring the achievement of goals which are more and more being stated in terms of social performance. The differences in the materials of instruction have been emphasized because of the two general approaches to this problem. One of these, the older, divided the social sciences into well-integrated parts: history, geography, economics, and civics. History, in turn, was divided into American, European, ancient, modern, and world. The second approach, the more recent, attempts to nucleate all the social sciences around dynamic problems of the present day.

In the first of these approaches the pupil studied certain bodies of knowledge which had been watered down from the college courses in the same area. The facts, gathered through research, were logically arranged and had a consistency and sequence all their own. Thus when a student had finished a course in American history he knew about the early settlements, the expansion of our country, the French and Indian Wars, the Revolutionary War, the formation of the Constitution, and so on down to the present. He had studied the facts, meaningful and otherwise, of American history. And so it was with the other histories, economics, geography, civics, etc.

The second of these approaches kept its eye on the present. It wanted the experience of mankind as presented in recorded history to be focused on the problems of the day. Many present problems could not be solved or really understood unless their history were known, their economics understood, or the nature of their geography comprehended. In this approach, the focus was not on history as such or on economics, but on the solution of the problem. How could students understand the problems of segregation of races without a knowledge of the history of slavery, the economic problems involved, the function of government in helping with the problem, and the question of the geographical distribution of races? This second approach is apt to omit much of

recorded history, a great deal of economic theory, and some of the intricacies of geography and to use only that material which helps solve the problem.

Because of these two somewhat contradictory methods of approach the selection of objectives toward which the teacher is working is doubly difficult.

OBJECTIVES IN THE TEACHING OF THE SOCIAL SCIENCES

The objectives selected from lists collected by workers in the field, from what teachers say they are striving to do, and from criticisms of tests will necessarily include a variety. Of the large number of objectives available, there will be included only those which are generally agreed to.

1. Information about social relations functional, meaningful facts.
2. Methods of acquiring information skills¹

- a. Read to understand
- b. Engage in group discussion
- c. Listen attentively to oral presentation of materials
- d. Consult maps to locate specific information
- e. Recite in class
- f. Read to locate information
- g. Consult charts and diagrams to locate specific items of information
- h. Make an outline or brief
- i. Give a special report or "floor talk"
- j. Make a summary or précis
- k. Draw a map
- l. Consult graphs and statistical tables to locate specific items of information
- m. Observe pictures, scenes, models, relics, exhibits, bulletin boards, etc., to locate specific items of information
- n. Write an expository theme explaining trend or point of view, a cause-effect relationship
- o. Read for enjoyment
- p. Read to memorize- intensive reading and rereading
- q. Take part in committee work

¹ The 20 objectives listed were taken from Kelley, T. L., and A. C. Krey, *Tests and Measurements in the Social Sciences*, pp. 64-69 (New York: Charles Scribner's Sons, 1934) by permission. Kelley and Krey gained the cooperation of 100 high school teachers in evaluating 52 items selected from a much larger number. The present list contains the 20 items which were rated highest, arranged in the order of their rating.

- r.* Observe pictures, scenes, models, relics, exhibits, bulletin boards, etc., for general impression and emotional enjoyment
- s.* Study maps to understand all the ideas they contain
- t.* Draw a diagram or chart

3. Evaluation of information

- a.* The ability to judge an event in the light of the times in which it occurs
- b.* The ability to weigh evidence and to judge the sources of information
- c.* The ability to comprehend causal relations
- d.* The ability to distinguish between relevant and irrelevant material

4. The development of attitudes and interests

- a.* The acquisition of desirable attitudes toward government, other races, other persons, standards of living and, in general, toward the problems of human relations
- b.* The development of some appreciation of the difficulties involved in the everyday problems of living
- c.* The acquisition of interest in good government, fair prices, the problems of capital and labor, conditions of work, and the good life.

When these objectives are considered as a whole, we find present information, skills and techniques of learning, judgment of the importance of sources of information, and appreciations, interests, and attitudes. A part of these objectives is concerned with actual participation in the process of socialization itself: taking part in committee work, making a report on some problem, reciting in class, or visiting a session of court or the legislature. Another part has to do with collecting and interpreting materials: learning to use tables of contents, indexes of books, standard reference works, encyclopedias, newspapers, maps, statistical tables, graphs, etc.

The third part, having to do with judgment, appreciations, attitudes, and interests does not come immediately from instruction but is an accumulation over the years as a result of good teaching and learning.

THE MEASUREMENT OF OBJECTIVES

It is apparent that these objectives differ in their ease of measurement. Easiest of all to measure is information, and for this reason, perhaps, many good measuring instruments of information have been developed. Our better tests are directed toward the meaning and significance of events and not toward facts as such. There is less emphasis,

for example, placed on the mere fact that the Magna Carta was signed in 1215 at Runnymede but more on this event as a milestone in the slow rise of individual freedom. We have also good tests of reading prose, reading and interpreting maps, and even to a lesser degree of the use of indexes, table of contents, and other reference materials. When we come to the motives involved in the participation in this socializing process or to attitudes and appreciations of the social scene, good standard tests simply do not exist. It can also be said that the newer problem type of instruction has up to the present been very difficult to measure. Most of our tests are still based on the older plan of dividing social science into history, economics, civics, and geography.

MEASUREMENTS OF ACHIEVEMENT IN THE SOCIAL STUDIES

ELEMENTARY SCHOOL

Test Batteries

In the elementary school social studies comprise history, civics, and geography. The tests for these three areas are usually *included in many general test batteries* under the caption "Social Studies." For the most part the tests of history and geography are kept separate. Two illustrations are now presented: (1) the Coordinated Scales of Attainment, and (2) the Metropolitan Achievement Tests.

The Coordinated Scales of Attainment have tests designed for use for each grade.¹ For this reason, a much wider sampling of significant facts in history and geography is available. The history test for grade 5 (Battery 5) consists of 60 items which cover the history of the United States through the Civil War. It has questions on the early discoverers, the struggles of the early colonists, the Declaration of Independence, the Revolutionary War, the formation of the Constitution, important inventions, the Civil War, etc. The preponderant emphasis is on the names of men whose accomplishments were outstanding. Two samples are:¹

22. Who organized the Committees of Correspondence?

- | | | | | |
|----------------------|---------------|-----------------|----------------------|------------------|
| 1. George Washington | 2. John Adams | 3. Samuel Adams | 4. Benjamin Franklin | 5. Patrick Henry |
|----------------------|---------------|-----------------|----------------------|------------------|

53. Who was the President of the United States during the Civil War?

- | | | | | |
|-----------|---------|-------------|------------|------------|
| 1. Wilson | 2. Polk | 3. McKinley | 4. Lincoln | 5. Madison |
|-----------|---------|-------------|------------|------------|

One can see in these illustrations the attempt to include functional questions.

In like manner there are 60-item history tests for each grade. The

¹ Items quoted by permission of Educational Test Bureau, Minneapolis, Minn.

following two items illustrate the type of questions used at the upper levels. The first is from Battery 7:

34. In the Northwest Ordinance, Congress set aside one section in each township for the support of
1. churches
 2. relief
 3. roads and bridges
 4. local government
 5. schools

The other is from Battery 8:

23. European nations were warned that they should keep out of American affairs by the
1. Ostend Manifest
 2. Monroe Doctrine
 3. Kentucky and Virginia Resolutions
 4. Hartford Convention
 5. Wilmot Proviso

The tests of geography suitable for grade 5 consist of 60 items. Eleven questions are based on the interpretation of two maps. There are questions on crops, imports, climate, products of states or countries, animals, harbors, cities, etc. Here again is the attempt to make the questions functional. The reason that a certain sort of wheat is called winter wheat, what latex is, how to calculate the shortest distance between two cities on a map, how ocean-going vessels reach Washington—these are typical questions. Two illustrations are:

49. A region in which the land, climate and vegetation are about the same is called a
1. unit region
 2. mountain region
 3. cultured region
 4. farming region
 5. natural region
31. The place in which iron is separated from the iron ore is called a
1. reducer
 2. separator
 3. blast furnace
 4. still
 5. purifier

Tests of geography suitable for each grade are furnished through grade 8. For example, Battery 7 (grade 7) begins with tests on the geography of Australia. There is a map of Australia together with six location questions. Altogether there are 15 questions about Australia—its animals, its wool, the location of its population, etc. The rest of the test has questions on Asia and Africa.

These tests of geography, which include many questions involving interpretation of facts and of maps, constitute very satisfactory tests which reflect the outcomes of instruction in the social studies.

The Metropolitan Achievement Tests¹ first introduce tests of social studies in the intermediate battery, designed for grades 5 and 6. Test 7 is labeled "Social Studies: History and Civics," and Test 8, "Social Studies: Geography." The test of history and civics is composed of 50 items, about one-fourth of which are on civics and the rest on history.

¹ Items quoted from this test by permission of World Book Company, Yonkers, N.Y.

In the items which test achievement in civics, children are asked to understand principles used in the selection of candidates, what city department first deals with criminals, and who immigrants and aliens are. The items on history include questions on the discoverers, on the Civil War and Southern recovery, on such inventors as Fulton, Stephenson, and Edison, and on colonization. The history and civics test which is contained in the advanced battery contains 51 items and is intended for grades 7 and 8, and the first half of grade 9. The questions on civics deal with problems of citizenship, immigrants, the regulation of airplane routes, and who usually does the picketing. There are 43 questions on history. Twelve of these questions are about persons such as Theodore Roosevelt, Booker T. Washington, Cabot, and Peary. There are questions about the Mexican War, Civil War, War of 1812, First World War, and Revolutionary War.

Illustrations from the intermediate battery are:

10. The purchase of Louisiana was important because:
 1. it didn't cost much 2. we bought it from France 3. it gave the United States complete control of the Mississippi Valley 4. it contained many Indians who wanted to trade furs for goods
44. In the main, police powers are exercised by the:
 1. states and local communities 2. Federal government 3. army
 4. criminologists

In general, the arrangement of these items seems to be without rhyme or reason. They certainly have no true chronological order. For example, a question about Egyptian civilization is followed by a question on carpetbaggers. As one contemplates such a mixture of items as appears in these tests he wonders if some other organization using more homogeneous groupings might not be more effective. It is this sequential arrangement which recommends the Coordinated Scales of Attainment.

The Metropolitan Achievement Tests also have both in their intermediate and advanced batteries of tests of geography. In the intermediate battery, Form R, there are 58 items in this test. There is a map of Florida and adjoining states which is used for questions about the interpretation of maps; questions are asked concerning the products, imports and exports, crops, industries, and occupations of a variety of states and countries. For the most part these questions refer to states or regions of the United States, but they extend to such places as Lapland, China, Germany, France, Mexico, Brazil, Iran, and Iraq. The student is asked to leap lightly from items such as

33. The principal racial element in Mexico is
 1. Negro 2. British West Indian 3. white 4. American Indian

to items such as

34. The chief export from Chile is

1. wool 2. meat 3. nitrate 4. coal

Geography in the advanced battery deals with topics similar to those in the intermediate battery. Its 53 items also ask questions about products, occupations, location of places, rivers and lakes, population, and industries of a great variety of states and countries. Questions are asked about Canada's most important natural resources, Java's leading products, leading occupation of the Chinese, and what Yugoslavia is expected to import. Many of the questions involve interpretation such as why the South can produce much cotton, why southeastern Alaska has developed more rapidly than other sections of Alaska, and what the smelting of iron ore requires.

The weakness of the use of judicious sampling in selecting test items for testing achievement in the social studies has been suggested. It does seem that, from the standpoints of curricular validity and of problem interpretation, items could be grouped around natural centers.

The Stanford Achievement Tests also have tests on social studies.

Specific Tests of the Social Studies

Many of the older tests of geography and history for the elementary school are now out of date. They are no longer valuable because their items are concerned too largely with small bits of information and consequently emphasize interpretation too little. For those interested a few are listed at the end of this chapter.

The Cooperative Social Studies Test for Grades 7, 8, 9 is one of the *newer* tests and one which undoubtedly is to be used in the upper grades. It is reviewed on page 195 of this text. Also worthy of consideration in this connection is the Kelty-Moore Test of Concepts in the Social Studies. There are two forms, of 35 concepts each, available for testing concepts acquired in the social studies from grade 4 through the junior high school.

Geography Tests

There are several geography tests suitable for testing in the elementary school. Two tests have been selected because they exemplify attempts to measure techniques and understandings gained from the study of geography rather than disconnected facts.

The Wiedefeld-Walther Geography Test is an illustration of a test that although old (1931) is still good because it was well built. It is divided into two parts, with three subheads under each part:

Part 1. Study abilities in geography

Test I. Reading

Test II. Organization

Test III. Map and graph reading

Part 2. Geography information

Test IV. Geography vocabulary

Test V. Geographical relationships

Test VI. Place geography

In spite of this test's many desirable characteristics, its usefulness has gradually disappeared and it is now out of print.

There is another test, too, which tries out the ability of pupils to interpret maps, graphs, charts, etc. This is Test B, work-study skills, of the Iowa Every-pupil Tests of Basic Skills. Two parts of Test B bear directly on the problem of measuring the outcomes of social studies.

Part I. Map reading—Sections A, B, and C

Part V. Reading graphs, charts, and tables

Part I has three sections, containing altogether 40 questions, with appropriate maps for each section. All maps are artificially constructed but include significant facts. An example with two questions is shown in Fig. 17.

Section C bases its 18 questions on eight maps which indicate elevation, temperature zones, cattle, chief railroads, rainfall, crop regions, principal mineral workings, and population. Each map represents the same hypothetical states. Using the data from these eight maps such questions as which state grows both tea and tobacco, which state probably leads in the production of hogs, and which state has the widest variety of minerals are asked. It is this use of functional, relational, and inferential questions which recommends this test to us so highly.

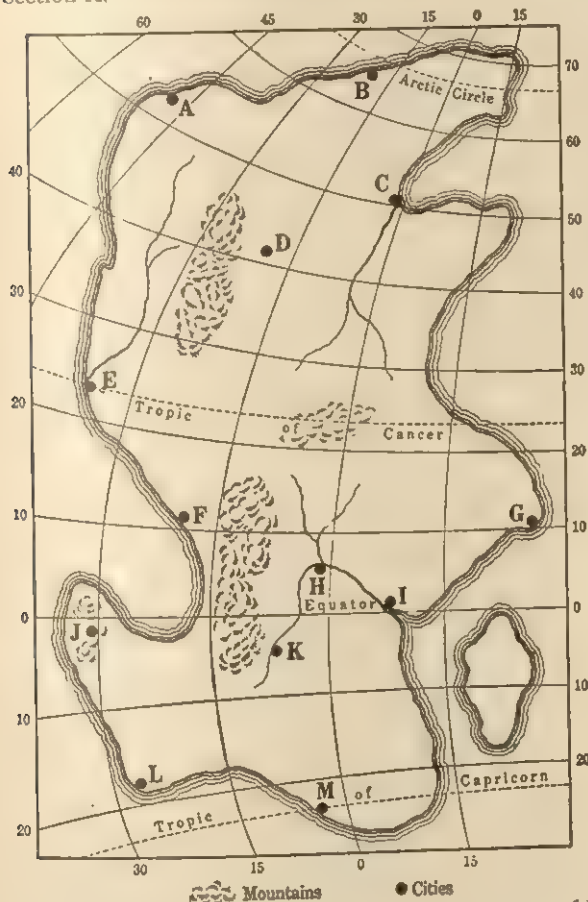
Part V on reading maps, charts, and tables bases all its questions on items selected from the social studies. Graphs based on the amount of merchandise exported, world production of automobiles, the average corn yield in Iowa and Pennsylvania, and sources of mechanical power in the United States, etc., are used to obtain answers to pertinent questions on these topics. There is also one table on tobacco yield and prices in eight states which forms the basis for the answer of four questions.

It is thus clear that, while this test on work-study skills is not intended as a test of geography, it contains adequate samples of the interpretation of maps and of graphs and tables composed of data drawn from the social studies. It is one test whose testing procedures might, and probably should, influence the teaching procedures of the

Part I. Map Reading

Section B

Directions: The questions to the right are based on the map below, which is a map of an imaginary land. Answer the questions in the same way that you did those in Section A.



16. Which of the following would be a probable cause of difficulty in building a railroad from K to I?

- 1) Lack of water
- 2) High mountains
- 3) Thick jungle
- 4) Lack of wood for ties

18. How does the longest day of the year at A compare with the longest day at L?

- 1) One cannot tell from the map
- 2) They are the same
- 3) The longest day at A is longer
- 4) The longest day at L is longer

FIG. 17. Work-study skills, Iowa Every-pupil Tests of Basic Skills. (By permission of Houghton Mifflin Company, Boston.)

social studies, for the techniques used test some of the most desirable outcomes of teaching.

SECONDARY SCHOOL

Testing Information and Meanings

Samples of tests taken from the various subjects which together constitute social science will first be illustrated and evaluated.

History Tests

Tests of American, European, ancient, and world history have been constructed. The most used of these are tests of American history.

The Cooperative American History Test is divided into two parts.¹ The first part, consisting of 62 multiple-choice items, samples events up to the Spanish-American War. The second part, of 36 items, samples events occurring between the end of the Spanish-American War and the time the test was constructed. While much of the test is purely factual, there is a definite attempt to present the questions in a functional, meaningful way. Illustrations from Form Q follow:

5. Large plantations were not established in the New England colonies chiefly because
 - 5-1 these colonies prohibited slavery
 - 5-2 most of the people lived in villages and towns
 - 5-3 nearly all capital was invested in commercial enterprises
 - 5-4 the soil and climate were not adapted to such a system
15. The outstanding hero of the Revolutionary War in the West was
 - 15-1 Nathanael Greene
 - 15-2 Ethan Allen
 - 15-3 George Rogers Clark
 - 15-4 Light-Horse Harry Lee
25. The Monroe Doctrine was intended to
 - 25-1 end our alliance with France
 - 25-2 prevent trade between Latin America and Europe
 - 25-3 promote American imperialism in the Caribbean
 - 25-4 prevent European interference in the Western Hemisphere

About one-half the questions cover the period between the first colonies and the Civil War. The first 12 items deal with pre-Revolutionary times. The norms are in terms of scaled scores in which the score of 50 represents the score "of an average child in average school with the usual amount of instruction." The scores run from 1 to 100 and are satisfactory bases of comparison both between tests and for the same pupil from one test to another.

Let us look more closely at these scaled scores used by so many of the Cooperative tests. A quotation from the Cooperative *Handbook* will throw more light on the meaning of scaled scores.

For example, the "50 point" on the *Cooperative American History Test* represents the score on this test made at the end of a year's typical instruction in the twelfth grade by a student having the following characteristics: (1) intelligence quotient in the seventh grade (where little selection has occurred) between 98 and 102;

¹ Items by permission of Educational Testing Service, Princeton, N.J.

(2) age between 14.25 and 14.75 as of grade 9.0; (3) score of 92 on the *New Stanford Achievement Test* at grade 8.4.

By assuming this group to be normal and to extend for 5 standard deviation units in either direction from the mean, there would be altogether 10 standard-deviation units. Now if we divide each of these units into 10 smaller parts, we have exactly McCall's T-score with a mean of 50 and an S.D. of 10. The 100 units along the base line are as nearly equal to each other as are any other units known to mental measurement.

A list of other tests of American History appears on page 203.

The Cooperative Modern European History Test satisfies more nearly the criteria for judging such a test than any other test of European history.¹ It is divided into two parts. Part I contains 62 items of the multiple-choice variety. It deals with the understanding of "fundamental movements and instructions as well as of personages, locations, and specific events." The second part attempts, without too much success, to measure historical judgment with 35 items. Perhaps a sample or two from each part will show something of the content and the manner of testing it. The following items are from Part I (Form Q):

10. One of the reasons that led Gustavus Adolphus to engage in the Thirty Years' War was
 - 10-1 the desire to recover territory lost after the death of Charles XII.
 - 10-2 the desire to aid the German Protestants.
 - 10-3 personal resentment against Cardinal Richelieu.
 - 10-4 fear of losing Norway.
- 24 The fall of the Bastille in 1789 was important because
 - 24-1 Lafayette was imprisoned there.
 - 24-2 large quantities of munitions and firearms were stored there.
 - 24-3 it was strategically located.
 - 24-4 it symbolized the tyranny of the government.

The following example is from Part II (Form Q):

13. The method proposed in the Covenant of the League of Nations for the prevention of war was the
 - 13-1 creation of a superstate with wide police powers.
 - 13-2 holding of an international plebiscite.
 - 13-3 establishment of compulsory arbitration.
 - 13-4 abolition of armaments.

The facts utilized in this test are well selected but with possibly too great an emphasis on political events and too little on economic and social ones. Its norms based on 6,000 cases from the high schools are

¹ Permission for using Cooperative Test items from Educational Testing Service, Princeton, N.J.

reported in scaled scores. The reliability of the test, .91 with a single grade, is satisfactory.

In tests similar to these tests just described the Cooperative Test Bureau has published satisfactory tests in ancient history and world history. The same principles of construction and standardization are used as were employed in the tests of American and European history.

Three other tests constructed by instructors at Kansas State Teachers College are worthy of consideration. These tests are called the Kansas American History Test, the Kansas Modern European History Test, and the Taylor-Schrammel World History Test.

Economics Tests

For those high schools which give a separate test in economics, the Cooperative Economics Test is available.¹ This test consists of two parts. Part I contains 15 items to be answered by matching statements and principles and by matching books and their authors, as well as 44 multiple-choice items. Part II contains 30 multiple-choice items. There is a wide sampling of the field of economics. Here is an illustrative sample of the matching items:

- | | | |
|-------------------------|--------------------------------------|-------|
| 1. Special assessments | 22. Largest source of revenue to the | |
| 2. Income tax | federal government | 22() |
| 3. Poll tax | 23. Largest source of revenue to | |
| 4. Sales tax | most local governments | 23() |
| 5. General property tax | 24. Can be easily shifted | 24() |

The following are two samples of the multiple-choice items, the first from Part I, and the second from Part II:

36. A factor tending toward inflation is
 - 36-1 an unbalanced federal budget.
 - 36-2 increased production of consumer goods.
 - 36-3 rising taxes.
 - 36-4 labor troubles.
27. Which term best describes the United Mine Workers of America?
 - 27-1 Trade union
 - 27-2 Industrial union
 - 27-3 Affiliated union
 - 27-4 Company union

The test consumes 40 minutes of testing time. It has percentile norms at the high school and college level and a reliability satisfactory for ordinary purposes. It has been criticized because Part I opens with Item 10 instead of Item 1 and Part II with Item 20. Furthermore, a few of the answers to the items might be challenged for their accuracy.

¹ Items by permission of Educational Testing Service, Princeton, N.J.

Civics' Tests

A few high schools stick to the older type of courses in civics and civil government, in which case the American Council Civics and Government Test might be helpful. This test is suitable for use in an advanced course in high school. It is divided into four parts. Part I is made up of 108 true-false items. Part II contains 13 matching exercises with five matches to be selected from eight possible ones. Part III contains 24 multiple-choice items, there being five choices for each item. Part IV is constructed of 23 completion items. All told, the test samples a wide area of the subject and uses 90 minutes of time. Its percentile norms are based on a rather small number of high school and college students. Its reliability is reported as .88 as computed by the Spearman-Brown formula. There are two forms of the test.

Testing Problems, Skills, and Procedures

The second method of approach to teaching of the social sciences emphasizes the focusing of facts upon the problems of the present. To secure a greater understanding of today's problems, emphasis must be placed on understanding of what is studied. To understand, the student must read with understanding, must be acquainted with the special terms embodied in reading, and in addition must know the techniques of reading graphs and tables and of discovering the sources of information. He must know when to use encyclopedias and atlases and how to take advantage of a table of contents or an index.

We shall describe and illustrate three tests as samples of what such tests do: (1) the Cooperative Social Studies Test for Grades 7, 8, 9, (2) the Cooperative General Achievement Test, Form X, and (3) Test of Critical Thinking in the Social Studies. Other tests do the same, but perhaps less well. It will be noticed that these include tests for the elementary school as well as the secondary school level.

The Cooperative Social Studies Test for Grades 7, 8, 9 is divided into three parts. Part I, Facts, Skills, and Applications, consists of 75 items, each with five choices in the answer.¹ It consumes 40 minutes of time in the taking. The items of the test cover a variety of subjects. One has to answer questions as to why the United States has decided at that time to build a larger navy, why Americans were more concerned about the First World War than about the Second World War, and for what gasoline taxes are most often used. It has problems on the interpretation of graphs and maps. A good illustration of the manner of test construction occurs in the following items (Form R):

¹ Educational Testing Service, Princeton, N.J. Items by permission.

44. Which one of the following has worked to the advantage of states in the poorer section of the country at the expense of the more prosperous states?
- 44-1 The workman's compensation law
 - 44-2 The federal relief system
 - 44-3 The wages and hours law
 - 44-4 Tariff on manufactured articles imported into the United States
 - 44-5 The method of collecting the income tax 44()
48. Which one of the following would be the best place to find an answer to the question: "What air line carried the most freight during 1939?"
- 48-1 An encyclopedia
 - 48-2 *The Reader's Guide*
 - 48-3 An atlas
 - 48-4 *The World Almanac*
 - 48-5 *Who's Who in America* 48()

Part II, Terms and Concepts, consists of 45 terms and concepts to be defined in 15 minutes. Such terms as "cabinet," "tenant farmer," "legislature," "revolt," "diplomat," and "levees" are illustrations.

24. Customs duties are collected when
- 24-1 goods are brought into a country.
 - 24-2 people pay an income tax.
 - 24-3 checks are cashed at a bank.
 - 24-4 a tax is collected for each article bought •
 - 24-5 people are fined for breaking a law 24()

Part III, Comprehension and Interpretation, is made up of seven passages of varying lengths about which questions are asked. The actual working time is 25 minutes. This part is a test of the understanding of reading passages from the social sciences. Percentile norms have been derived.

Another test much like this but geared to the level of the high school is the Cooperative Test of Social Studies Abilities which parallels the present test in Parts I, II, and III but adds a fourth part called Applying Generalizations.

These two tests make it possible to evaluate the outcomes of instruction in a manner different from using the amount of information acquired for this purpose. The test can easily be used as a diagnostic device.

The Cooperative General Achievement Tests is one of the more recent (1947) of the Cooperative test series. It is divided into two parts. In Part I, Terms and Concepts, 15 minutes is the time allotted for identifying the correct definitions of 50 terms. The student is asked to know the meaning of "the Black Death," "depreciation," "filibuster," "enfranchised," "plutocracy," and "federation." He is called upon to know the principal results of the Crusades, what the advocates

of a short ballot want, and what an agrarian economy is. Part II, Comprehension and Interpretation, is pretty largely a reading test employing seven short paragraphs and one graph about which questions are asked. The whole test, which takes 25 minutes of working time, is to discover the ability to read and interpret such material.

The third selection, Test of Critical Thinking in the Social Studies by J. Wayne Wrightstone, is divided into three parts, each of which consumes 15 minutes in taking. In the elementary series, meant for grades 4 to 6, Part I itself is divided into three sections which altogether ask 36 questions. The first section of Part I furnishes tables of prices, of the production of hogs, and of population, location, principal products of towns, graphs of production and altitude. Questions are then asked directly on these data. The second division consists of six questions on the location of facts. The third division is on the capacity to use an index. Part II, on drawing conclusions from facts, is more distinctive than any other part. The instructions themselves indicate immediately a different sort of test:

Mark with: (+) every statement which is true and can be proved by the facts stated.

(0) every statement which might be true but cannot be proved by the facts stated.

(-) every statement which is false as shown by the facts stated.

Here is an example from the test:¹

III. When bricks are taken out of the kiln or oven they are red and very hard. They are ready for use. Bricks will last for hundreds of years. They will not decay and fall to pieces as wood does. They will not burn. They are not costly. These qualities make bricks very useful in building and they often take the place of wood.

- | | |
|--|-------|
| 9. Bricks vary in price, quality and make | 9() |
| 10. Bricks have many good qualities which sometimes make them more useful than wood | 10() |
| 11. Bricks have many more lasting qualities than wood | 11() |
| 12. Because bricks spoil rather quickly and are so expensive, they cannot take the place of wood | 12() |

Part III, on applying general facts, consists of nine paragraphs with a matching test for each paragraph. The directions and one sample will indicate the procedure used.

Directions: This section has a number of paragraphs. Below each paragraph are two sets of statements about the paragraph. In the left hand column are five statements. Three of those statements will help you to understand the three refer-

¹ By permission of Bureau of Publications, Teachers College, Columbia University, New York, and of J. Wayne Wrightstone.

ences in the right hand column. Select a statement from the left hand column which best explains a reference in the right hand column. Write the number of the statement in the space after the reference.

VI. Although new traffic rules are being made all the time, there are still many automobile accidents. Every year thousands of people are killed by careless drivers. Hundreds of children are killed while playing in the streets. Although there seem to be too many automobiles, they are very useful in business and transportation. The building of elevated roads and the invention of new safety devices would help reduce accidents.

- | | |
|---|---|
| 1. Most laws are made to help the people. | 16. Explains why traffic rules are set up..... () |
| 2. Machines have helped us make greater progress. | 17. Explains why automobiles are important in industry..... () |
| 3. Improvement of machines needs an inventive people. | 18. Explains how new safety devices may help reduce accidents () |
| 4. Transportation follows natural roads. | |
| 5. Industry in these days needs science. | |

This Test of Critical Thinking in the Social Studies deserves special consideration because it attempts to measure one of the most important outcomes of social instruction, *i.e.*, critical thinking. Such thinking usually involves a consideration of the facts which have already been collected, comparison of the facts both among themselves and with others, and a judgment rendered as to the quality of the facts or as to the conclusion drawn. One will note that the critic is not the creator who discovers and solves the problem. The critic renders value judgments about the materials already produced. If these criteria for critical thinking are correct little, if any, critical thinking is needed in the part concerned with obtaining facts. Part I of this test is simply a test of work skill involved in obtaining facts from tables, graphs, etc.

Part II, on drawing conclusions, and Part III, on applying general facts, contain much that would fall under the category of critical thinking. Consider the illustration about bricks referred to in a previous paragraph. The requirement that the pupil make a judgment about a statement which may be drawn from the facts partakes of the nature of critical thinking. Furthermore, the rendering of a judgment of negation adds strength to the exercise. In the illustration about traffic rules the pupil compares two statements and decides whether one is explained by, or is an illustration of, the other. This may not be critical thinking at the adult level, but at the fifth- or sixth-grade level there is no doubt that it partakes of the nature of critical thinking.

The results from this test correlate highly with scores on Modern

School Achievement Test and the New Stanford Achievement Test as well as with McCall's Multi-mental Scale, a scale of intelligence. These facts indicate that perhaps critical thinking enters into the taking of all tests and plays a large part in reading. They also imply that perhaps this test of critical thinking is nothing new, after all, but another test of the skills demanded in the mastery of the materials of social studies. The test has satisfactory reliability and a manual which offers excellent instructional procedures to use with those pupils who have low scores on the test.

Tests of Social Terms

The understanding of written material in the area of social science may be measured (1) by the number of questions asked about a paragraph or selection, or (2) by selecting those terms that are characteristic of treatment of social relations and making a test for them. Among the tests of social terms are (1) the Wesley Test in Political Terms, (2) the Wesley Test of Social Terms, (3) Pressey's Test of Concepts Used in the Social Studies, and (4) the Kelty-Moore Test of Concepts in the Social Studies.

The Wesley Test in Political Terms is composed of items which are functional and which have wide applicability. The test terms were selected from the Krey-Kelley list of 4,000 words and terms used in the social sciences. The separate items were evaluated by 27 college instructors and 13 members of the working staff. Political terms included those with military, diplomatic, and legal implications and other terms which are related to government. After considerable experimentation the final test was cast in the best-answer type and has four forms of 10 items each. The reliability is .68 for each part but when all 40 words are used the reliability is satisfactory, for individual diagnosis. The Wesley Test in Social Terms differs from the Wesley Test in Political Terms (1) in selection of items, and (2) in length. This test includes items from all the social studies instead of from one area alone. There are 80 items for each form. The correlation of each of these tests with intelligence tests, with reading and with tests of civics indicates that while these tests are somewhat related to all of them they also measure something quite different. Samples of terms measured in the test of social terms are "smuggling," "sheriff," "selectman," "scab," "remonstrance," "regimented," "public utility," "proxy," "propaganda," "proclamation," "penalty," "paternalism," "notary," and "monarchist." Its use of the best-answer form of construction implies that all the answers are partially correct and the one is to be selected which most nearly answers the question. "The form allows the use of false, but attractive, associations, of partially correct ideas, and of a variety of

options based upon similarities of sound and form."¹ These two tests were made originally for grade 12 but are applicable from grade 9 to sophomores in college. The Kelty-Moore Test of Concepts in the Social Studies is intended for younger children. Forms X and Y, containing 56 items for tryout in each grade, were prepared for grades 4 through the junior high school. These items were submitted to "three leading authorities on the teaching of social studies" and the tests tried out experimentally on 100 fourth-grade pupils, 100 sixth-grade pupils, and 100 eighth-grade pupils. From these procedures, 70 items were selected and divided into two forms of 35 items each. The reliability of these forms was about .77. When all 70 items are merged into one test the reliability is raised to .85 or .90.

The third test, constructed by L. C. Pressey, is called Test of Concepts Used in the Social Sciences.² The terms for this test were selected from some 1,444 words which had been collected from a variety of sources and then evaluated by 64 high school teachers, 5 professors of college history, and 7 individuals specially trained to be sensitive to the sociological value of each word outside the history classroom. Thus from the number of times the term occurred, from the judgment of history teachers, and from their sociological value 346 items were selected and tested. The terms were arranged in Forms A, B, C, and D containing 85 items in Form A and 80 each for Forms B, C, and D. These are not parallel forms in the ordinary statistical sense but rather forms made from items selected more or less at random. As illustrations from the test, here are two items from Form A:³

16. Which word refers to the affairs relating to one's own country?
 (a) foreign (b) international (c) domestic (d) diplomatic
41. What happens when money depreciates?
 (a) it becomes less valuable (b) it will buy more (c) it has to go back to the mint (d) it can be used in foreign countries

The following two items are from Form B:

34. Which word refers to political corruption?
 (a) graft (b) lynching (c) revolt (d) mutiny
66. What is the outer edge of a civilized area called?
 (a) metropolis (b) suburbs (c) frontier (d) seacoast

Because children grow up in a reading and talking world they acquire habits of seeing and reading words whose meanings may be vague and at times incorrect. Studies of children's vocabularies have made progressive teachers very sensitive to this inadequacy on the part of the

¹ Kelley and Krey, *op. cit.*, p. 222 (an article by Edgar B. Wesley).

² Published in Kelley and Krey, *op. cit.*

³ Items used by permission of Luella Cole.

pupils. It is for this reason that these tests of terms and phrases are so important. While some of the tests described are not as well standardized as we should like, they represent a movement in the right direction. These tests should be supplemented by teacher-made tests of terms, so that what is learned about social interaction may be clear and well understood.

Measurement of Attitudes in the Social Sciences

In Chap. 17 appears a discussion of the formation and measurement of attitudes. The present treatment presupposes what is there presented and offers a sample of attempts to measure some of those attitudes which are supposed to grow directly out of courses in social sciences. Since attitudes are so instrumental in determining the action which is taken, they need great clarity in definition and precise instruments to measure their attainment. Unfortunately, neither of these outcomes has been satisfactorily achieved.

The usual attitude test or scale consists of a series of statements with which the subject may express agreement or disagreement. Such a scale is the Wrightstone Scale of Civic Beliefs which¹ is suitable for grades 9 to 12. This scale is divided into four parts:

| Part | Statements |
|--|------------|
| I. Racial attitudes..... | 20 |
| II. International attitudes..... | 20 |
| III. National political attitudes..... | 20 |
| IV. Attitudes toward national achievements and ideals..... | 20 |

After each statement there is an A and a D. The directions say, "If you agree with the statement, make a heavy black mark in the space under A. If you disagree, make a mark under D. Be sure to mark every statement, and use a question mark only in extreme cases of doubt." Illustrations are selected from each part. The first two are from Part I:

- | | | |
|--|---|---|
| 5. The white race is no better nor worse than other races. | A | D |
| 12. The United States should prohibit Chinese immigration. | A | D |

The next two are from Part II:

- | | | |
|---|---|---|
| 26. Most of our immigrants are undesirables from other nations. | A | D |
| 35. The United States should pursue a liberal policy towards immigration. | A | D |

The next two are from Part III:

- | | | |
|---|---|---|
| 46. Only a traitor refuses to fight for his country. | A | D |
| 52. Business and industry increasingly need some government regulation. | A | D |

¹ World Book Company, Yonkers, N.Y. Items used by permission.

Finally, here are two from Part IV:

65. Most criminals tend to be feeble-minded and ignorant.

A D

75. Only radicals and socialists join labor unions.

A D

The scale is easily scored and furnishes percentile norms for grades 9, 10, 11, and 12. The percentiles thus obtained indicate the amount of liberalism or conservatism which an individual possesses. For example, if a subject receives a percentile score of 75, this means that the individual is more liberal than 75 per cent of the standardized group and less liberal than 25 per cent.

The test was validated by using only items which were present and in common use in textbooks. The items used were checked by 21 social scientists as to whether the agreement or disagreement was interpreted as being liberal or conservative. On the results of their judgment answers to items are scored as progressive or conservative. The reliability of the test is indicated by a coefficient of .94.

Weakness in arriving at a satisfactory measurement of attitudes by means of this procedure stems from the method itself. Subjects may not know what the statement means, they may have generalized from too limited experiences, and they may prevaricate. Each of these is discussed in Chap. 17, "The Measurement of Attitudes."

SUMMARY

Two general approaches to the problems involved in teaching social science have caused the measuring instruments to be strikingly different. On the one hand we have tests of history, economics, sociology, and geography; on the other, there are tests of the techniques used in acquiring information, of critical thinking, and of interests and attitudes. The best tests of history sample functional, meaningful material. They use the test forms of multiple choice, completion, and matching. Answers are sometimes so arranged that any one of the answers might be the correct one but the *best* answer is the one desired. In critical thinking comparisons are made between statements and judgments are inferred from the data presented. Doubt is implied as to whether certain inferences could or could not be drawn from the data presented. Instruments were also presented by means of which attitudes could be registered.

In general, it was found that the objectives in teaching social sciences are very numerous and that many of them have not as yet been satisfactorily measured. Among these latter are interests, social participation both in school affairs and in after-school life, and attitudes. Satisfactory objective tests of a student's ability to marshal his information about a

single topic and arrange it in a convincing manner have not as yet been constructed. Despite the tendency of many tests of the special subjects to emphasize the acquisition of information as such, it still may be truthfully averred that there are many useful standardized tests in the social sciences.

LIST OF TESTS IN SOCIAL SCIENCE

I. HISTORY

American History

1. Cooperative American History Test, high school. 1933-1940. Forms T, X, and Y. Time: 40 minutes. Authors: Howard R. Anderson, E. F. Lindquist, Charlotte W. Croon, and Harry Berg. Cooperative Test Service, New York.

2. Kansas American History Test, high school and college. Two forms. Two levels. Time: 40 minutes. Authors: Arthur Hartung, H. E. Schrammel, and C. Stewart. Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kans.

3. Coordinated Scales of Attainment in American History, grades 7-8. 1932-1933. One form. Time: 45 minutes. Authors: Mary G. Kely and M. J. Van Wagenen. Educational Test Bureau, Minneapolis, Minn.

4. American History Test, National Achievement Test, grades 7-8. 1937-1939. Two forms. Nontimed (about 50 minutes). Authors: Robert K. Speer, Lester D. Crow, and Samuel Smith. Acorn Publishing Co., Rockville Center, N.Y.

5. Test of Factual Relations in American History, grades 10-12. 1936. Two forms. Nontimed (about 100 minutes). Author: Eugene S. Farley. Educational Test Bureau, Minneapolis, Minn.

6. Cray American History Test, high school. 1951: Reliability: .87-.91. Factual information, 28 items; skills, 16 items; interpretation of historical information, 8 items; understanding of historical processes, 26 items; reasoned inferences, 12 items. World Book Company, Yonkers, N.Y.

World History

1. Cooperative World History Test, high school. 1934-1937. Forms X and Y. Time: 90 minutes. Authors: H. R. Anderson and E. F. Lindquist. Cooperative Test Service, New York.

2. Taylor-Schrammel World History Test, high school. 1936. Test I, first semester; Test II, second semester. Time: 40 minutes. Authors: Wallace Taylor and H. E. Schrammel. Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kans.

3. Iowa Academic Contest, Every-pupil Tests, high school. New forms each year. World history. Bureau of Educational Research and Service, University of Iowa, Iowa City.

4. Cooperative Contemporary Affairs Test of High School Classes. 1940. One form for each year. Time: 120 minutes. Authors (1940 edition): Alvin C. Eurich, Elmo C. Wilson, Edward A. Krug, *et al.* Cooperative Test Service, New York.

5. Iowa Academic Contest, Every-pupil Tests, High School Contemporary Affairs. Bureau of Educational Research and Service, University of Iowa, Iowa City.

European History

1. Cooperative Modern European History, high school and college. 1937-1940. Forms N, O, P, and Q. Time: 40 minutes. Authors: H. R. Anderson, Wallace Taylor, E. F. Lindquist, Charlotte W. Croon, and Mary Willis. Cooperative Test Service, New York.

2. Kansas Modern European History, Test II, high school. 1938. One form. Time: 40 minutes. Authors:

Alvin L. Hasenbank and H. E. Schrammel. Kansas State Teachers College, Emporia, Kans.

3. American Council European History, grades 10-15. 1929. Two forms. Time: 90 minutes. Authors: Harry J. Carman, Walter C. Langsam, and Ben D. Wood. World Book Company, Yonkers, N.Y.

4. Vannest Diagnostic Test in Modern European History, high school. Bureau of Cooperative Research, Indiana University.

Ancient History

1. Cooperative Test in Ancient History, high school. 1938-1939. Forms O and P. Time: 40 minutes. Authors: Howard R. Anderson, E. F. Lindquist, Wallace Taylor, and Charlotte W. Croon, *et al.* Cooperative Test Service, New York.

II. CIVICS AND GOVERNMENT

1. Cooperative Test in American Government, high school. Forms X and Y. Time: 40 minutes. Author: John Haefner. Graphic and verbal material, functional and interpretive. Cooperative Test Service, New York.

2. Cooperative Test of Community Affairs, high school, Form 4. Key made to fit individual community. Time; 30 minutes. Authors: Ray A. Price and Robert F. Steadman. Cooperative Test Service, New York.

3. American Council Civics and Government Test, high school and college. 1929. Two forms. Time: 90 minutes. Authors: Rober D. Leigh, Joseph D. McGoldrick, Peter H. Odegard, and Ben D. Wood. Reliability: .88. World Book Company, Yonkers, N.Y.

4. Iowa Academic Contest, Every-pupil Tests, American Government, high school. Bureau of Educational Research and Service, University of Iowa, Iowa City.

5. Mordy-Schrammel Elementary Civics Test, elementary grades and

junior high school. Kansas State Teachers College, Emporia, Kans.

6. Hill Test in Civic Attitudes, grades 6-12, Public School Publishing Company, Bloomington, Ill.

7. Hill Test in Civic Information, grades 6-12. Public School Publishing Company, Bloomington, Ill.

8. Hill-Wilson Test in Civic Action, grades 6-12. Public School Publishing Company, Bloomington, Ill.

III. ECONOMICS

1. Cooperative Economics Test, high school and college. 1939. Forms P and S. Time: 40 minutes. Authors: Howard R. Anderson, J. E. Partington, *et al.* Cooperative Test Service, New York.

2. American Council Economics Test, high school and college. World Book Company, Yonkers, N.Y.

3. Iowa Academic Contest, Every-pupil Tests, high school economics. Bureau of Education Research and Service, University of Iowa, Iowa City.

IV. SOCIOLOGY

1. Black-Schrammel Sociology Test, high-school and college. Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kans.

V. GEOGRAPHY

1. Wiedefeld-Walther Geography Test, grades 4-8. 1931. Four forms. Time: 60 minutes. Authors: N. Theresa Wiedefeld and E. Curt Walther. World Book Company, Yonkers, N.Y.

2. Brueckner-Cutright Practice Exercises in Locational Geography, elementary grades and junior high school. Educational Test Bureau, Minneapolis, Minn.

VI. SOCIAL SCIENCE

1. Test of General Proficiency in the Field of Social Studies, high school. Cooperative General Achievement Tests, revised series. Forms Q and R. Time: 40 minutes. Authors: Mary Willis *et al.* Cooperative Test Service, New York.

2. Cooperative Test of Social Studies Abilities, high school. 1916-1939. Form Q. Time: 80 minutes. Authors: J. Wayne Wrightstone *et al.* Cooperative Test Service, New York.

3. Test of Critical Thinking in the Social Studies, grades 4-6. 1938-1939. Two forms. Time: 45 minutes. Author: J. Wayne Wrightstone. Bureau of Publications, Teachers College, Columbia University, New York.

4. Kansas Social Studies Unit Tests, grades 4, 6, and 8. Kansas State Teachers College, Emporia, Kans.

5. Kelty-Moore Tests of Concepts in the Social Studies, grades 4-9. Authors:

M. G. Kelty and N. E. Moore. Charles Scribner's Sons, New York.

6. Wesley Test in Social Terms, grades 6-16. 1932. Two forms. Nontimed (about 30 minutes). Author: Edgar B. Wesley. Charles Scribner's Sons, New York.

7. Wesley Test in Political Terms, high school. Charles Scribner's Sons, New York.

8. Kepner Background Test of Social Studies in High School. Ginn & Company, Boston.

9. Pressey Tests of Concepts Used in the Social Studies, high school. 1934. Charles Scribner's Sons, New York.

QUESTIONS AND EXERCISES

1. Distinguish sharply between the two points of view which give direction to the teaching of social science.

2. Explain four types of objectives formulated for the teaching of social science.

3. Distinguish between *inert* facts and *functional* facts.

4. *a.* Name and explain four objectives for which there are no satisfactory objective tests.

b. Secure copies of the Metropolitan Achievement Tests and the Coordinated Scales of Attainment. Make a careful comparison of the (1) items, (2) sampling of historical facts, and (3) general value of their two history tests.

5. Devise a rating scheme for measuring the participation of students in club activities.

6. How would you arrive at a judg-

ment of children's interests in social-science activities?

7. Critically evaluate the use of the multiple-choice type of question in measuring the outcomes of history teaching.

8. Explain the meaning of a scaled score. What are its uses?

9. What are some of the outcomes tested by the Cooperative Social Studies Test?

10. Do you think such a test as that of Wrightstone really tests *critical thinking*? Why?

11. Describe Wrightstone's Attitude Scales. Enumerate its strong points and its weak ones.

12. Which of the tests mentioned measures the capacity to read in the social sciences?

BIBLIOGRAPHY

Books

BUROS, OSCAR K. (ed.): *The Nineteen Forty Mental Measurements Yearbook*, Items 1614-1642. Highland Park, N.J.: The Mental Measurements Yearbook, 1941.

—: *The Third Mental Measurements Yearbook*, Items 590-619. New Brunswick, N.J.: Rutgers University Press, 1949.

The Forty-fifth Yearbook of the National Society for the Study of Education, Part I, "The Measurement of Understanding," Chap. V. Chicago: University of Chicago Press, 1946.

GREENE, HARRY A., ALBERT N. JORGENSEN, and J. RAYMOND GERBERICH: *Measurement and Evaluation in the Secondary School*, Chap. XVII. New York: Longmans, Green & Co. Inc., 1943.

KELLEY, T. L., and KREY, A. C.: *Tests and Measurements in the Social Sciences*, pp. 1-119, 153-233, 234-339. New York: Charles Scribner's Sons, 1934.

SMITH, EUGENE R., RALPH TYLER, *et al.* *Appraising and Recording Student Progress*, Chap. III. New York, Harper & Brothers, 1942.

TOWNSEND, AGATHA: "The Reliability and Validity of the USAFI American History Test," in 1947 *Achievement Testing Program in Independent Schools and Supplementary Studies*, pp. 53-58, Educational Records Bulletin No. 48. New York: Educational Records Bureau, 1947.

TRAXLER, ARTHUR E.: *Techniques of Guidance*, pp. 90-93. New York: Harper & Brothers, 1945.

WESLEY, EDGAR BRUCE: *Teaching the Social Studies*, Chap. XXIII. Boston: D. C. Heath and Company, 1937.

Articles

"How Do Senior College Students and Adult Groups Stand on the 'Times' Test?" *School and Society* (1943) 57:654.

LINDQUIST, E. F.: "The Form of the American History Examinations of the Cooperative Test Service," *Educational Record* (1931) 12:459-475.

PRICE, ROY A., and ROBERT F. STEADMAN: Part 8, "Testing for Community Information," pp. 213-225, "Utilization of Community Resources in the Social Studies," *Ninth Yearbook of the National Society for Social Studies*, 1938.

READ, JAMES MORGAN: "History versus the Social Sciences," *School and Society* (1943) 58:149-151.

TRAXLER, ARTHUR E.: "Progressive Methods as Related to Knowledge of American History," *School and Society* (1943) 57:640-643.

CHAPTER 8

Measurement of Foreign Languages

OBJECTIVES IN TEACHING

The objectives customarily sought in the teaching of any foreign language may be classified under four heads:

1. A knowledge of the language itself. This involves the ability to read, write, spell, and speak the language. The materials used for mastering this language may vary from newspapers and magazines written in this foreign tongue to selections from its classics. It involves the mastery of vocabulary, verb forms, idioms, agreements among words, inflections, and other minutiae which are needed for reading, speaking, and understanding the language.

2. An appreciation of the literature written in that language. Even in elementary courses some acquaintance is achieved with the masterpieces which express realistically and artistically the great experiences of mankind.

3. An appreciation of the geography, history, manners, customs, and culture of the foreign country whose language is being studied. Some years ago Nicholas Murray Butler, then President of Columbia University, spoke of teachers of the foreign languages as the ambassadors who represented foreign countries and who helped students become acquainted with the fine points of their civilizations. They were not to think of themselves as teachers of a language only.

4. Interrelations between that language and English. English has borrowed from many foreign languages. Thorndike's studies showed that 52 per cent of ordinary running words are derived from the Latin and another 11 per cent from the Greek through the Latin. Many phrases have been adopted unchanged into English. English grammar, too, sometimes has its principles more sharply focused when contrasted with that of a foreign language. If the teacher keeps this objective clearly in mind and strives to achieve it, considerable improvement in the knowledge of the derivation of English words and a better understanding of the structure of English can be achieved.

The history of testing itself shows attempts to measure many of these objectives. The Columbia Research Bureau and the American Council

on Education have constructed a variety of French tests on reading, grammar, and vocabulary. The Columbia Research Bureau has constructed an Aural French Test while Lundeborg and Tharp have an Audition Test in French. In the area of history, manners, and customs at least one test, Miller's, French Life and Culture, has been constructed. Trabue, too, made a scale to aid in measuring French composition.

THE MORE MEASURABLE OBJECTIVES

As time went on and data accumulated on these tests of language achievement, it became increasingly clear that the facts involved in learning the language itself were more susceptible to accurate measurement than the other less well defined and less well agreed-upon areas. At any rate, a careful study of the most successful language tests at the present time indicates that they attempt to measure the following specific objectives:

1. Reading with understanding
2. Vocabulary growth
3. Knowledge of functional grammar
4. Translation into English and vice versa

Many teachers wish for a standardized test of conversation and pronunciation. They also want their pupils to know about history, manners, and customs of the people. Above all, they would like some measuring instrument for the influence on English of the study of foreign language. After a brief consideration of the best available tests in French, Spanish, German, and Latin, the author will evaluate them.

TESTS OF FRENCH

Those who construct the best French tests today must make their tests conform to facts and principles which research has made available.¹ A vocabulary test, for example, must select its words from the words most frequently used. Such a test uses suitable words from Vander Beke's *French Word Book* or the word lists of Henmon and Cheydleur. The former book, regarded by many critics as the best, contains 6,136 words selected because of their frequency of use in thousands of French running words. The questions on grammar and usage must be functional and must be selected from those generally proved to be the minimum essentials for understanding written language. And finally, the selections for reading must be long enough to

¹ Publications of the American and Canadian Committees on Modern Foreign Languages contain excellent research materials on many aspects of test construction (see Bibliography).

develop rather thoroughly one idea and must be arranged in steps of increasing difficulty.

Because the Cooperative Test Series of the American Council on Education utilized to the best advantage principles based on research, they are generally regarded as the leading language tests today. A study of the 16 double-column pages of critical evaluation of French tests in the *Nineteen Forty Mental Measurements Yearbook* (Buros) showed that the tests of the American Council received fewer criticisms and far more commendations than any other tests. Such expressions appeared as "the first place among educational measurements today," "only praise for the grammar test," and "a better measuring instrument than the traditional examination of yesterday." Not all statements are as flattering as are these, but the general trend is highly favorable.

The cooperative tests are issued each year so that new techniques and criticisms can be embodied in the latest forms. These yearly editions make it possible for the new test to embody changes that take place in the curriculum.

A good illustration of the cooperative test series appears in the Cooperative French Test,¹ revised series, elementary, Form O. This test has three parts: reading (15 minutes), vocabulary (10 minutes), and grammar (15 minutes). Scores may be had for each of the parts and for the test as a whole.

The *reading* part has 40 items. Each item consists of a statement in French which is followed by five choices from which the correct answer is to be selected. The following illustrations are from Form O:

9. Les hommes qui composent une armée sont
 - 9-1 tous des officiers.
 - 9-2 des avocats.
 - 9-3 des militaires.
 - 9-4 des paysans.
 - 9-5 des invalides.
14. On appelle la pièce ou l'édifice où l'on trouve beaucoup de livres
 - 14-1 la cuisine.
 - 14-2 la chambre à coucher.
 - 14-3 la bibliothèque.
 - 14-4 le pupitre.
 - 14-5 le corridor.

The *vocabulary* test presents its choices for the correct answer in English. There are 50 words varying in difficulty from *chaud*, *tout*, and *pied* through *pluie*, *bâtiment*, and *profond* to *papillon*, *lorsque*, and *honteux*. Each item is presented as follows:

¹ Items of test by permission of Educational Testing Service, Princeton, N.J.

19. fois
 19-1 faith
 19-2 time
 19-3 hour
 19-4 sausage
 19-5 flower
18. désespérer
 18-1 disturb
 18-2 descend
 18-3 despair
 18-4 deserve
 18-5 describe

The *grammar* part has 35 items largely concerned with usage such as plurals, idioms, agreement of pronominal adjective and noun, pronouns, indirect object, verbs that use *être* or *avoir*, past participles, etc. The answers are in French.

26. Are you cold?
 (_____) froid?
 26-1 Avez-vous
 26-2 Faites-vous
 26-3 Étiez-vous
 26-4 Faisiez-vous
 26-5 Êtes-vous
28. He left immediately.
 Il (_____) parti immédiatement.
 28-1 a
 28-2 est
 28-3 était
 28-4 avait
 28-5 faisait

Tables are furnished whereby scaled scores can be transmuted immediately into percentiles computed for the end of the semester. Norms based on (1) public secondary schools of the East, Middle West, and West and on (2) public secondary schools of New England are available. The reliability of this test has been variously reported as .93 to .97.

LIST OF FRENCH TESTS

I. GENERAL

1. Cooperative French Test. Elementary form, 1-3 semesters in high school; advanced form, 2 years high school. Authors: elementary form, Jacob Green-

berg and Geraldine Spaulding; advanced form, Geraldine Spaulding and Paul Vaillant. Time: 40 minutes. Cooperative Test Service, New York.

2. American Council Alpha French

Test, grades 9-16. Two parts. Two forms. Part I, vocabulary and grammar; Part II, silent reading and composition. Time: 40 minutes. World Book Company, Yonkers, N.Y.

3. American Council Beta French Test, grades 7-11. Two forms. Part I, vocabulary; Part II, comprehension; Part III, grammar. Time: 90-100 minutes. World Book Company, Yonkers, N.Y.

4. American Council French Grammar Test, grades 9-16. Two forms: Time: 22-27 minutes. World Book Company, Yonkers, N.Y.

5. American Council on Education French Reading Test, 2 semesters or more of college French. Time: 50 minutes. World Book Company, Yonkers, N.Y.

6. Columbia Research Bureau French Test, grades 9-15. Time: 90 minutes. World Book Company, Yonkers, N.Y.

II. AURAL

1. Columbia Research Bureau Aural French Test, grades 9-16. Two forms. Time: 45-60 minutes. World Book Company, Yonkers, N.Y.

2. Lundeborg-Tharp Audition Test in French, high school and college. Two forms. James B. Tharp, Ohio State University, Columbus, Ohio.

III. OTHER TESTS

1. French Life and Culture, high school and college. One form. Time: 40 minutes. Author: Minnie M. Miller. Bureau of Educational Measurements, Kansas State Teachers College. Emporia, Kan.

2. French Reading, grade 10. Two forms. Time: 30 minutes. Department

of Educational Research, University of Toronto.

3. French Vocabulary Test, grades 9-10. Two forms. Time: 30 minutes. Department of Educational Research, University of Toronto.

4. Standard French Test, high school. Vocabulary, grammar, and comprehension. One form. Time: Part I, 28 minutes; Part II, 32 minutes. Public School Publishing Company, Bloomington, Ill.

5. Cooperative French Test, lower and higher levels. Lower level, 1-2 years high school; higher level, more than 2 years in high school. 1942-1947. Forms S and X. Time: 80-85 minutes. Authors: Geraldine Spaulding, Laura Towne, and Sarah Woolfson Lorge. Cooperative Test Service, New York.

6. Examination in French Grammar, high school. Lower level, 1944, 1-2 years in high school, Form LFG-1-B-4; upper level, 1945, 2½ years in high school, Form UFG-1-B-4. Time: 40-45 minutes. Authors: Examinations Staff of the U.S. Armed Forces Institute. Cooperative Test Service, New York.

7. Examination in French Reading Comprehension, high school. Lower level, 1944, 1-2 years high school, Form LFR-1-B-4; upper level, 1945, 2½ years high school, Form UFR-1-B-4. Time: 50-55 minutes. Authors: Examinations Staff of the U.S. Armed Forces Institute. Cooperative Test Service, New York.

8. Examination in French Vocabulary, high school. Lower level, 1944, 1-2 years in high school, Form LFV-1-B-4; upper level, 1945, 2½ years in high school, Form UFV-1-B-4. Time: 40-45 minutes. Authors: Examinations Staff of the U.S. Armed Forces Institute. Cooperative Test Service, New York.

SPANISH TESTS

From the many tests of Spanish, the author has selected only those prepared by the Cooperative Test Service and by the Columbia Research Bureau.

The Cooperative Spanish Tests are prepared after the manner of

their French tests. The Cooperative Spanish Test, junior form, is divided into three parts:

| Part | Time, minutes |
|---------------------|---------------|
| I. Reading..... | 15 |
| II. Vocabulary..... | 10 |
| III. Grammar..... | 15 |

The reading test consists of 40 sentences and short paragraphs which are answered in Spanish (junior form).¹

39. A los discípulos que no son listos es difícil
- 39-1 castigarlos
 - 39-2 enseñarles
 - 39-3 aprenderlos
 - 39-4 encontrarlos
 - 39-5 mirarlos
22. Los hombres que viven mucho tiempo llegan a ser
- 22-1 conocidos
 - 22-2 largos
 - 22-3 verdes
 - 22-4 ancianos
 - 22-5 jóvenes

Part II contains 50 Spanish words ranging from easy to hard. The definitions are in English.

3. comprender
- 3-1 understand
 - 3-2 buy
 - 3-3 eat
 - 3-4 take away
 - 3-5 promise
22. triste
- 22-1 road
 - 22-2 truthful
 - 22-3 trunk
 - 22-4 suit
 - 22-5 sad
30. paso
- 30-1 price
 - 30-2 paste
 - 30-3 part
 - 30-4 paving
 - 30-5 step

¹ Items of test by permission of Educational Testing Service, Princeton, N.J.

- 36. ciego
 - 36-1 sky
 - 36-2 seal
 - 36-3 continuous
 - 36-4 blind
 - 36-5 wax

Part III, on grammar, has 35 items. Each item has a statement in English followed by translation into Spanish except for the omission of a crucial word which illustrates the point of grammar.

- 16. It is half past six.
 - () las seis y media.
 - 16-1 Es
 - 16-2 Está
 - 16-3 Son
 - 16-4 Hay
 - 16-5 Están
- 10. He has lost his books.
 - Ha perdido () libros.
 - 10-1 suyos
 - 10-2 suya
 - 10-3 de él
 - 10-4 su
 - 10-5 sus
- 61. They have just opened it.
 - () de abrirlo.
 - 61-1 Acaban
 - 61-2 Hubieron
 - 61-3 Tenían
 - 61-4 Han
 - 61-5 Están

The same strong points are present in this test as were present in the French test. The reliability is high ($r = .95$). There are many forms of the test, and percentile norms are prepared for public secondary schools in the South, the East, Middle West, and the West along with norms for independent secondary schools and colleges. The Cooperative Spanish Test, revised series, advanced, uses the same amount of time for each of the tests as does the elementary test. The material in every case is more advanced.

LIST OF SPANISH TESTS

- 1. Cooperative Spanish Test, revised series, elementary form. Part I, reading, 15 minutes; Part II, vocabulary, 10 minutes; Part III, grammar, 15 minutes. Percentile norms for high school and college students. Forms N, O, and

P. Time: 40 minutes. Reliability: .95 (odds versus evens). Authors: Jacob Greenberg, Robert H. Williams, and Geraldine Spaulding. Cooperative Test Service, New York.

2. Cooperative Spanish Test, revised series, advanced form. Part I, reading, 15 minutes; Part II, vocabulary, 10 minutes; Part III, grammar, 15 minutes. Percentile norms for high school and college. Forms N, O, P, and Q. Time: 40 minutes. Reliability: .98 (odds versus evens). Authors: E. Herman Hespelt, Robert H. Williams, and Geraldine Spaulding. Cooperative Test Service, New York.

3. Columbia Research Bureau Spanish Test, high school and college. 1926-1927. Forms A and B. Part I, vocabulary, 25 minutes; Part II, comprehension, 20 minutes; Part III, grammar, 45 minutes. Time: 90 minutes. Reliability: .97. P.E.-*meas.* = 3. Authors: Frank Callcott and Ben D. Wood. World Book Company, Yonkers, N.Y.

4. Examination in Spanish Grammar, lower level, 1-2 years of high school or 1 year of college. 1944. Form B. Time: 40-45 minutes. Separate answer sheets must be used. Authors: Examinations Staff of the U.S. Armed Forces Institute. Cooperative Test Service, New York.

5. Examination in Spanish Reading Comprehension, lower level, 1-2 years of high school or 1 year college. 1944. Form B. Time: 40-45 minutes. Must use separate answer sheets. Authors: Examinations Staff of the U.S. Armed Forces Institute. Cooperative Test Service, New York.

6. Examination in Spanish Vocabulary, lower level, 1-2 years of high school Spanish or 1 year in college. 1944. Form B. Time: 40-45 minutes. Must use separate answer sheets. Authors: Examinations Staff of the U.S. Armed Forces. Cooperative Test Service, New York.

7. Lundeberg-Tharp Audition Test in Spanish, high school and college. 1944. Form B. Time: 30 minutes. Authors: Olav K. Lundeberg and James B. Tharp. James B. Tharp, College of Education, Ohio State University, Columbus, Ohio.

8. Iowa Placement Examinations, Spanish Training, Series S.T. revised, grades 12-13. 1924-1926. Forms A and B. Time: 43(50) minutes. Authors: C. E. Seashore, G. M. Ruch, G. E. Vander Beke, and G. D. Stoddard. Bureau of Educational Research and Service, State University of Iowa, Iowa City, Iowa.

GERMAN TESTS

Tests of German are constructed in the same manner as those of French and Spanish.

The Cooperative German Test, revised series, elementary Form N has also three parts:

| Part | Time, minutes |
|---------------------|---------------|
| I. Reading..... | 15 |
| II. Vocabulary..... | 10 |
| III. Grammar..... | 15 |

The test of reading consists of 40 sentences, the answers to which are in German. The following illustrations are from Form N:¹

¹ Items of test by permission of Educational Testing Service, Princeton, N.J.

17. Um frische Luft ins Zimmer
zu lassen, öffne ich
17-1 den Ofen
17-2 den Schrank
17-3 das Buch
17-4 den Mund
17-5 das Fenster
2. In der Klasse sehen wir die
Schüler und
2-1 den Schneider
2-2 den Arzt
2-3 den Lehrer
2-4 den Kaufmann
2-5 den Fleischer
13. Unser Wohnzimmer ist
13-1 auf der Strasse
13-2 in dem Garten
13-3 in der Schule
13-4 in unserem Haus
13-5 im Hospital
12. Es ist zwölf Uhr mittags.
Wir sollten jetzt
12-1 schlafen gehen
12-2 frühstücken
12-3 zu Abend essen
12-4 zu Bett gehen
12-5 zu Mittag

In Part II, on vocabulary, there are 50 words to be defined in English.

17. manchmal
17-1 alternate
17-2 sometimes
17-3 on time
17-4 certain
17-5 each one
18. froh
18-1 rough
18-2 frothy
18-3 happy
18-4 reddish
18-5 furrowed
24. Stimme
24-1 monster
24-2 obstinacy
24-3 voice

- 24-4 forehead
- 24-5 stimulant
- 43. Sammlung
 - 43-1 sample
 - 43-2 similarity
 - 43-3 appliance
 - 43-4 collection
 - 43-5 foundling

Part III, on grammar, contains 35 items. Each item has first an English sentence and then the German translation with a significant word omitted. This answer is found among five German words.

- 13. An hour has sixty minutes.
() Stunde hat sechzig Minuten.
 - 13-1 Einem
 - 13-2 Eine
 - 13-3 Einer
 - 13-4 Ein
 - 13-5 Einen
- 11. The beautiful lady is my aunt.
Die () Dame ist meine Tante.
 - 11-1 schön
 - 11-2 schöne
 - 11-3 schönen
 - 11-4 schöner
 - 11-5 schönes
- 5. Now I speak only English.
Jetzt () ich nur Englisch.
 - 5-1 spricht
 - 5-2 sprecht
 - 5-3 sprach
 - 5-4 spreche
 - 5-5 sprich

The advanced form uses the same divisions and takes the same time, but the questions and problems are at a more advanced level. There are a number of forms of the test and the percentile norms are prepared for secondary schools of the North, East, and West but not the South. The reliability is satisfactory ($r = .95$ or $.96$). The correlation with teachers' marks varies from $.65$ to $.69$.

LIST OF GERMAN TESTS

- | | |
|--|---|
| 1. Cooperative German Test, Elementary Form, grades 6-9, 1-6 semesters. Revised series. Forms N, O, and P. Part I, reading, 15 minutes; Part II, | vocabulary, 10 minutes; Part III, grammar, 15 minutes. Reliability: $.95$. Cooperative Test Service, New York. |
| | 2. Cooperative German Test, Ad- |

vanced Form, 4 semesters or more. 1938-1940. Forms N₁/O, P, and Q. Time: 40 minutes. Reliability: .96. Cooperative Test Service, New York.

3. American Council Alpha German Test, grades 9-16. 1926-1927. Two forms, two parts. Part I, vocabulary and grammar; Part II, Silent Reading and Composition. Time: 40 minutes. World Book Company, Yonkers, N.Y.

4. Columbia Research Bureau German Test, grades 9-15. 1926-1927. Two forms. Part I, vocabulary, 25 minutes; Part II, comprehension, 20 minutes; Part III, grammar, 45 minutes. World Book Company, Yonkers, N.Y.

5. American Council on Education German Reading Test, 2 semesters or more of college German. 1937-1938. Forms A and B. Time: 50 minutes. World Book Company, Yonkers, N.Y.

6. Examination in German Grammar, lower level, high school and college, 1-2 years of high school. 1945. Form B. Must use separate answer sheets. Time:

60-65 minutes. Authors: Examinations Staff of the U.S. Armed Forces Institute. Cooperative Test Service, New York.

7. Examination in German Reading Comprehension, lower level, high school and college, 1-2 years. 1945. Form B. Must use separate answer sheets. Time: 50-55 minutes. Authors: Examinations Staff of the U.S. Armed Forces Institute. Cooperative Test Service, New York.

8. Examination in German Vocabulary, lower level, high school and college, 1-2 years. 1945. Form B. Must use separate answer sheets. Time: 45-50 minutes. Authors: Examinations Staff of the U.S. Armed Forces Institute. Cooperative Test Service, New York.

9. Lundeborg-Tharp Audition Test in German, high school and college. 1929. Forms A and B. Authors: Olav K. Lundeborg and James B. Tharp. James B. Tharp, College of Education, Ohio State University, Columbus.

ITALIAN TESTS

Suitable tests for Italian have been constructed under the leadership of the Cooperative Test Service.

LATIN TESTS

The Latin tests of the Cooperative Achievement Tests also concentrate on reading, vocabulary, and grammar. The teaching objectives of Latin teachers are well measured in these tests. One Latin prognostic test is included which has great possibilities as an instrument of guidance.

The Cooperative Latin Test, revised series, elementary, form Q has three parts:¹

| Part | Time, minutes |
|---------------------|---------------|
| I. Reading..... | 15 |
| II. Vocabulary..... | 10 |
| III. Grammar..... | 15 |

A total score may also be computed. In the reading test there are two types of items. In the first 11 items a sentence is written in Latin with an essential word or phrase omitted. The omitted word or phrase which is correct appears among four other words or phrases which are

¹ Items of test by permission of Educational Testing Service, Princeton, N.J.

incorrect. Here are some illustrations from elementary Form IX, experimental:

10. Servus bonus in agris (- -).
 - 10-1 armābit
 - 10-2 laborābit
 - 10-3 timēbit
 - 10-4 movēbit
 - 10-5 portābit
13. Quid in bellō timētis? (- -) timēmus.
 - 13-1 periculum
 - 13-2 agrum
 - 13-3 puerōs
 - 13-4 flūmen
 - 13-5 oculōs

The remainder of this part consists of three paragraphs in Latin with questions in English. Three questions are asked about each paragraph.

Part II consists of 50 Latin words to be defined in English. The words range from easy to hard. Three items from elementary Form P will illustrate the type of word and the form of the test:

8. trēs
 - 8-1 three
 - 8-2 tree
 - 8-3 very
 - 8-4 effort
 - 8-5 sad
27. inferō
 - 27-1 flee
 - 27-2 yield
 - 27-3 interfere
 - 27-4 compare
 - 27-5 bring into
42. jam
 - 42-1 since
 - 42-2 for
 - 42-3 already
 - 42-4 though
 - 42-5 before

Part III, on grammar, has 35 items. Each item consists of a sentence in English, its translation save for one word or phrase, and then five choices among which the correct answer is found. The cases of nouns, tenses of verbs, agreement of noun and adjective, uses of the ablative

dative cases, and so on are included. Two illustrations are taken from elementary Form P:

7. I gave the queen a horse.

(. —) dedī.

7 1 Rēginam equum

7 2 Rēginae equum

7-3 Rēginae equō

7 4 Rēginam equō

7-5 Rēgina equum

19. We were in the camp.

(——) erāmus.

19-1 In castra

19-2 In castram

19-3 In castris

19-4 Castris

19-5 In castrās

The advanced form of the Cooperative Latin Test has more complex sentences. The paragraphs to be read, the words to be defined, and the grammar are distinctly more difficult than those of the elementary form.

Percentile norms are available for these tests both for high school and colleges. As for the other Cooperative Achievement Tests, separate norms are furnished for public secondary schools and for independent secondary schools. The reliability of these tests is reported from .94 to .96.

Because many students find Latin so difficult and make such little progress in mastering the language it is often a moot question as to whether some students should take it. Two measuring instruments are of aid here. The first of these, any good intelligence test, has already been discussed. Such a test correlates markedly with achievement in the course. The second measuring instrument is called a prognostic test. The Orleans Solomon Latin Prognostic Test presents a controlled situation in the learning of Latin. Seven actual lessons are learned and applied in a defined amount of time. It is assumed that the progress made in this miniature preview will be an earnest of future success. The actual correlation with subsequent achievement as measured by a combination of teachers' marks and achievement tests was reported by the authors to be .80. If a student were low in intelligence and low on this prognostic test his chances of learning Latin successfully would be small indeed.

Two other prognostic tests have been constructed for foreign languages. The Foreign Language Prognosis Test by Percival M. Symonds and the Luria-Orleans Modern Language Prognosis Test. The former of

these, suitable for use in grades 8 or 9, has two forms and correlates .60 and .61 with achievement-test scores. Its working time is 44 minutes. The Modern Language Prognosis Test by Max A. Luria and Jacob S. Orleans claims to measure the ability of students to learn Spanish, French, or even Italian. It can be used from grade 7 through grade 13. The test requires 76 minutes to take. The correlation of .68 between prognostic-test scores and scores on achievement has not been found by other investigators. Kaulfers,¹ for example, found correlations ranging from .35 to .52 between prognostic-test scores and achievement-test scores or teachers' marks.

It would thus appear that some prognostic tests of modern foreign language have not proved to be very effective in predicting subsequent standings in the language in question. One must remember that with a correlation of .60 a test's forecasting efficiency is only 20 per cent better than chance. Prognostic tests, however, can be used along with many other factors as confirmatory evidence for or against taking up the study of a foreign language.

LIST OF LATIN TESTS

1. Cooperative Latin Test, elementary form, revised series. First 3 semesters of high school and college. Forms N, O, P, Q, and R. Reading, 15 minutes; vocabulary 10 minutes; grammar 15 minutes; also total score. Percentile norms for high school and college. Reliability: .96. Author: George A. Land. Cooperative Test Service, New York.

2. Cooperative Latin Test, advanced form, revised series, High school and college. Forms P, Q, and R. Reading, 15 minutes; vocabulary, 10 minutes; grammar, 15 minutes; also total score. Percentile norms for high school and college. Reliability: .94. Correlation with school marks and total score: .76-.81. Correlation of test scores and regent examinations: .71. Author: Forms Q and R, George A. Land; Form P, John C. Kirtland. Cooperative Test Service, New York.

3. Orleans-Solomon Latin Prognosis Test, high school and college. 1926. Seven lessons in Latin, include knowl-

edge of masculine and feminine, use of cases, vocabulary, verb forms, translation, English derivatives, singular and plural. Correlation of this test and average of teachers' marks and achievement tests: .80. Authors: Jacob S. Orleans and Michael Solomon. World Book Co., Yonkers, N. Y.

4. A. Cooperative Latin Test, lower level, high school and first 2 years of college; higher level, more than 2 years in high school. 1942. Form S. Time: 80 minutes. Authors: Harold V. King and Geraldine Spaulding. Norms: tentative scaled scores. Cooperative Test Service, New York.

5. Kansas First Year Latin Test, high school, first and second semesters. 1936. Two forms. Two levels. Test 1, Forms A and B, first semester; test 2, Forms C and D, second semester. Time: 40(45) minutes. Authors: Mary Alice Seller, Lois Bellinger, and H. E. Schrammel. Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kans.

¹ Kaulfers, Walter Vincent, *The Forecasting Efficiency of Current Bases for Prognosis in Junior High School Beginning Spanish*, unpublished doctor's thesis, Stanford University, 1933.

EVALUATION OF TESTS OF FOREIGN LANGUAGES

Many of the criticisms leveled at French and other foreign-language tests at an earlier date have been met. The selection of words for the vocabulary tests has been improved, errors of fact have been eliminated, and questions have been arranged in the order of difficulty. The criticism that New England and New York norms might not be suitable for the rest of the country has been met by constructing norms for the public secondary schools of the South, of the East, West, and Middle West and for the independent secondary schools of New England. Present-day criticism revolves around (1) the test forms, (2) the content of our best tests, and (3) the omissions.

Present-day evaluation of foreign-language tests is concerned about the very form of the objective test itself. The critics hold that the process of recognition of the correct answer out of five alternatives is a passive process quite different from actual recall of a word in a translation situation. Moreover, in such definitions of words only one meaning is used, while the essence of language rests in the variety of meanings a word can convey according to the context. A quotation here from Henmon answers this objection: "The reply is that the recognition method gives more pupil response in the same length of time, that scoring is easier and more objective, and that while the absolute scores by the completion or recall method are considerably lower, the correlation between results of this with those attained by the recognition method are almost as high as the reliabilities of either technique."¹ Evidence for this last statement is furnished in German vocabulary tests in which the reliabilities varied from .89 to .94 and the correlations between a completion test and a five-response recognition test varied from .81 to .87.

Foreign-language tests are subject to other shortcomings. It is claimed, for example, that the selections for translation are entirely too short and are not unified. Critics are also fearful that the presentation of four *wrong* answers may affect the students' learning, for they should hear and see only the correct forms. Other critics are sure that verb forms are inadequately sampled, or that pronouns get only a cavalier treatment. They are fearful, too, that since the tests have to do with the learning of the structure and meaning of the written language, teaching will be strongly influenced in the same direction.

The second group of evaluators are not satisfied with what the tests contain. They bewail the omission of tests of conversation and pronun-

¹ Henmon, V. A. C., *Achievement Tests in the Modern Foreign Languages*, Publications of the American and Canadian Committees on Modern Languages, Vol. V, p. 10. New York: The Macmillan Company, 1929.

ciation. They think, too, that certainly something should be done about testing the transfer effect to the vernacular from the foreign language. They think, for example, that the Miller test, French Life and Culture, should be improved and brought up to date. Perhaps such knowledge might provide a tremendous motivating influence on the learning of the language itself.

Those who make the tests admit many of these contentions but feel that until the teachers themselves agree on some constructive program the building of satisfactory standardized tests is well-nigh impossible. If the teacher himself is carefully trained in test construction such as is developed in Chap. 3, then he can make sound tests for individual objectives. One attempt to measure aural French by the Columbia Research Bureau was not too successful because the pronunciation of teachers differed widely and too much of the French was based on written rather than on conversational French.

SUMMARY

Of the four leading objectives (1) ability to read, write, speak, and understand the language; (2) ability to appreciate its literature; (3) ability to appreciate the geography, history, and the manners and customs of the countries speaking the language; and (4) the interrelations between that language and English—only the first has been successfully measured with standard tests. Attempts have been made to measure these other outcomes but with indifferent success.

The constructors of the Cooperative Achievement Tests have taken advantage of the criticisms leveled at the earlier tests and of the tremendous amount of available research and have constructed highly reliable, valid, and well-standardized tests in the foreign-language area. They have narrowed their test to the testing of three areas: (1) reading, (2) vocabulary, and (3) grammar. Cooperative tests in French, Spanish, German, and Latin have been presented and illustrated. It was shown that all these tests had high reliability, an abundance of forms, separate tests for elementary and advanced students, and percentile norms for both high school and college. These tests even go so far as to present norms for different types of secondary schools and for schools located in different areas of the United States.

In spite of these excellencies many thoughtful teachers think that the form in which the test is constructed, tests only the capacity to recognize the right answer, a mental process very different from a translation. Some of them think that the presentation to the student of wrong answers may have a bad effect; while others emphasize the importance of aural tests.

It was pointed out that many of the desirable objectives in the teach-

ing of foreign languages have had as yet no satisfactory standardized tests constructed.

QUESTIONS AND EXERCISES

1. Describe the four objectives usually striven for in the teaching of any foreign language. Which one of these has proved most susceptible to measurement? Why?

2. What did President Butler imply by calling teachers of foreign languages "ambassadors"?

3. What features are usually included in a good French test? How reliable is it? How valid?

4. What sources of information of a research nature are available for test constructors in French?

5. Is the selection of the meaning of a French word from five alternatives the same as translating it? What was the evidence offered by Professor Henmon

bearing on this point? Do you think that Henmon's evidence answered the question?

6. If a person can translate a short passage well, can he also translate a long passage well?

7. Why is it difficult to construct a satisfactory aural test? One of life and culture?

8. Compare the French tests with the German and Spanish ones. Are there any differences in test construction?

9. What are the salient characteristics of a prognostic test? Describe one such test.

10. What are the means available for advising a student about taking Latin?

BIBLIOGRAPHY

BUROS, OSCAR KRISEN (ed.): *The Nineteen Forty Mental Measurements Yearbook*, Items 1340-1375. Highland Park, N.J.: The Mental Measurements Yearbook, 1941.

———: *The Third Mental Measurements Yearbook*, Items 178-213. New Brunswick, N.J.: Rutgers University Press, 1949.

GREENE, HARRY A., ALBERT N. JORGENSEN, and J. RAYMOND GERBERICH: *Measurement and Evaluation in the Secondary School*, Chap. XVI, New York: Longmans Green & Co., 1943.

Handbooks of The Cooperative Achievement Tests. New York: Cooperative Test Service.

HAWKES, HERBERT E., E. F. LINDQUIST, and C. R. MANN (eds.): *The Construction and Use of Achievement Examinations*, Chap. VI. Boston: Houghton Mifflin Company, 1936.

ODELL, C. W.: *Educational Measurements in High School*, Chap. VI, New York: Appleton-Century-Crofts, Inc., 1940.

PETERS, EMMA: "Relation of Tests to Improvement of Instruction," *Classical Journal* (1932) 28:187-196.

Publications of the American and Canadian Committees on Modern Languages. New York: The Macmillan Company, 1929.

BUCHANAN, MILTON A.: *A Graded Spanish Workbook*, Vol. III.

CHEYDLEUR, F. D.: *French Idiom List*, Vol. XVI.

HAUCH, EDWARD F.: *German Idiom List*, Vol. X.

HENMON, V. A. C.: *Achievement Tests in the Modern Foreign Languages*, Vol. V.

KENISTON, HAYWARD: *Spanish Idiom List*, Vol. XI.

MORGAN, B. Q.: *German Frequency Workbook*, Vol. IX.

VANDER BEKE, GEORGE E.: *French Work Book*, Vol. XV.

RUCH, G. M., and GEORGE D. STODDARD: *Tests and Measurements in High School Instruction*, Chap. VIII. Yonkers,

N.Y.: World Book Company, 1927.

SEIBERT, LOUISE C., and EUNICE R. GODDARD: "The Use of Achievement Tests in Sectioning Students," *Modern Language Journal* (1934) 18:289-298.

SYMONDS, P. M.: *Measurement in Secondary Education*, Chap. VIII. New York: The Macmillan Company, 1927.

———: "A Foreign Language Prog-

nostic Test," *Teachers College Record* (1930) 31:540-556.

TRAXLER, ARTHUR E.: *Techniques of Guidance*, pp. 81-84. New York: Harper & Brothers, 1945.

WRIGHTSTONE, J. WAYNE: "Measuring Diverse Objectives and Achievement in Latin," *Classical Journal* (1938) 34:155-165.

CHAPTER 9

Measurement of Mathematics

IMPORTANCE OF MATHEMATICS IN OUR MODERN WORLD

At no time in the history of the world has the importance of quantity, timing, and precision been more clearly demonstrated and more fully recognized than during the Second World War and since that time. Mathematics is the indispensable tool of precision in measures involving quantity and time. The natural sciences owe most of their progress to the use of measurement. Their slogan has been "Unless a thing is measured its nature remains unknown." But probably the most dramatic applications of mathematics in recent years have occurred in the areas of the social sciences. The outstanding tool in the quantification of the social sciences has been statistics. Furthermore, even betting odds are now calculated with mathematical nicety. Mathematics, then, justifies its place in school as an introduction to science and scientific thinking as well as in the workaday activities of trade and commerce.

TESTS OF MATHEMATICS IN THE ELEMENTARY SCHOOL

OBJECTIVES IN TEACHING ARITHMETIC

In the broadest sense, the objective in teaching arithmetic is to aid pupils to appreciate and understand the quantitative aspects of daily life. It involves the capacity to use our number system in making more precise measurements of all kinds, in innumerable transactions involving money and the interchange of goods, in the calculations of time and distance, in the construction of objects of all kinds, and in many other situations. To accomplish this broad aim more specific objectives are necessary:

1. To acquire an understanding of the vocabulary used in quantitative thinking. In addition to the language of quantity such symbols as equal, square root, and degrees, minutes, and seconds must be learned. This means the capacity to translate written descriptions of quantitative transactions into accurate computations with numbers.

2. To learn to perform quickly and accurately the four fundamental operations of addition, subtraction, multiplication, and division with whole and mixed numbers, common and decimal fractions, and denominate numbers.

3. To gain a deeper and more precise understanding of business transactions involving such problems as interest on money, discount, bonds, commissions and profits, taxation, school finances, banking, etc., by translating general statements about them into ideas involving quantity.

4. To acquire the ability to solve problems that are described in words or that arise in ordinary living. In some cases this involves the collections of facts bearing on a problem, the analysis of the problem, the decision about the operation or operations to use, and the correct manipulation of the processes involved.

5. To learn to understand the quantitative aspects of problems arising in everyday living so that judgments about them will be more precise. Among these problems the advantages of judicious spending and saving and of thrift are of great importance.

Another list of objectives has been summarized in the following statements which grew out of the construction of the Cooperative Mathematics Test for Grades 7, 8, and 9. A committee first drew up a list of 12 objectives for mathematics at this level. "These objectives may be seen to fall into four general categories corresponding to the four parts of the test: (1) mathematical skills, (2) mathematical facts, terms, and concepts, (3) mathematical applications, and (4) appreciation of the nature and value of mathematics."¹

SURVEY TESTS FOR USE IN THE ELEMENTARY SCHOOL

All general test batteries for the elementary school have sections on both the fundamentals and the problems of arithmetic. Because habits in arithmetic are arranged in an increasingly complex manner as learning progresses, the coverage especially of the fundamentals is in most cases quite satisfactory. Here are two samples: (1) the Metropolitan Achievement Tests, and (2) the California Achievement Tests.

The Metropolitan Achievement Tests, intermediate battery (grades 4, 5, and 6), has one section on arithmetic fundamentals and one on arithmetic problems. The section on arithmetic fundamentals includes the addition, subtraction, multiplication, and division of whole numbers, common fractions, and decimal fractions. In each of these operations, the examples begin with the simplest operations such as adding two single numbers, proceed to the addition of nine single numbers in a column, and continue to the addition of five five-place numbers. Zero difficulties appear at suitable points. Fractions are complicated both by

¹ *Manual of the Cooperative Test Service*. The committee was composed of Alice H. Darnell, Rose E. Lutz, Stevenson W. Fletcher, Jr. and John C. Flanagan. By permission,

introducing mixed numbers and by requiring the subtraction of fractions with different denominators. Decimal fractions increase in difficulty to such examples as .003).0156. A few of the 60 examples deal with percentage and a few with the addition and subtraction of denominate numbers. The section on problems in arithmetic deals with a variety of written problems, only a few of which have grown out of the actual experiences of children. A few samples of problems which might grow out of a child's experiences are (1) the calculation of the number of boxes that would be needed if a girl has 255 candles and puts five in a box; (2) the calculation of Sol's earning at 40 cents an hour if he works from 8:30 to 11:00 and from 2:30 to 3:30; and (3) the distance club members can walk between 8:30 and noon if they walk $2\frac{1}{2}$ miles an hour. Sample problems are concerned with the computation of the monthly income if the total yearly income is known, of the average monthly cost of gas if you know what the total cost per year is, and of the number of feet of wire fencing needed if the dimensions of a field are known. There are 40 problems. If we check the Metropolitan Arithmetic Test against the aims and objectives in teaching arithmetic we find a good coverage of most of the objectives described. There is no separate section for the testing of the vocabulary and symbols used in arithmetic. The problems, too, are more influenced by adult needs than by the experiences of children. There is no special technique already worked out for purposes of diagnosis. A teacher, though, may obtain considerable understanding of a child's weaknesses by an analysis of his paper.

One other illustration, the California Achievement Tests, will be given. This test specializes on the fundamentals. It omits social science, science, and literature and therefore can give a much more complete treatment of these fundamentals. It is divided into four levels: primary (grades 1, 2, 3, and 4), elementary (grades 4, 5, and 6), intermediate (grades 7, 8, and 9), advanced (high school and college). For our purposes we shall describe the elementary and intermediate batteries.

The elementary battery (grades 4, 5, and 6) is made up of seven sections. Its first two sections have 30 items concerned with the meaning of words and symbols used in arithmetic. It asks what "two hundred six" indicates in numbers, or "one thousand two." It samples the meaning of Roman numerals, and asks about the smallest of four numbers. It asks about the meaning of $+$, $-$, \times , and \div , $\%$, lb., $\sqrt{\quad}$, etc. Then follows a set of increasingly difficult problems which grow largely out of the experiences of children. Following these problems are one whole page of additions, one of subtraction, one of multiplication, and one of division. Except for the arrangement, which is very convenient

for studying each child's strong points and difficulties, the manipulations required differ very little from those of the Metropolitan Achievement Test. Altogether there are 105 items dealing with arithmetic while the Metropolitan has 100. The Metropolitan tests use 40 problems; the California tests, 15. The California tests include a plan already worked out and keyed for the analysis of difficulties. The intermediate battery resembles the elementary battery in form but differs in the following ways: the terms and written numbers are more difficult, for example, three-eighths, DCC, and " $a \frac{5}{6} b \frac{3}{4} c \frac{7}{8} d \frac{2}{3}$ find the largest number." The symbols to be known include the greatest common divisor as well as the formulas for measuring the volume of a prism and the area of a triangle. There are four pages of fundamentals as in the elementary battery. Opportunity also exists for analyzing errors by keyed references. This test does offer the teacher an opportunity for analyzing a child's results. The first two parts are tests of mathematical words and symbols. These two improvements make this a strong test for measuring arithmetic.

Other batteries which contain good tests of arithmetic are (1) the Stanford Achievement Test, and (2) the Coordinated Scales of Attainment.

SEPARATE TESTS FOR ARITHMETIC

More complete tests entirely devoted to mathematics are also available.

The Cooperative Mathematics Test for Grades 7, 8, and 9 attempts to measure the four objectives described on page 226. It is divided as shown in the accompanying table. The section on skills, Form Q, con-

| | Time, minutes | Reliability | N, eighth grade |
|------------------------------------|------------------|-------------|--------------------|
| I. Skills..... | 30 | .88 | 170 |
| II. Facts, Terms and Concepts..... | 10 | .69 | 170 |
| III. Applications..... | 30 | .86 | 170 |
| IV. Appreciation..... | 10 | .72 | 170 |
| Total..... | 80 | .92 | |

tains addition, subtraction, multiplication, and division of whole numbers, mixed numbers, common fractions, decimal fractions, percentage, mensuration, and solution of simple formulas which involves some elementary algebra. Part II asks questions about the meaning of range in a series of numbers, altitude of a triangle, discount, diameter of a circle,

hexagon, hypotenuse, meter, etc. In Part III mathematical applications are made to percentage of school children promoted, miles on a speedometer, cost of gas, table of contents of a book, percentage of bone in meat, thickness of ice and number of people allowed to skate, etc. Part IV deals with the recognition of facts missing from a problem.¹

2 A motorist used 10 gallons of gasoline on a trip. How many miles per gallon did he average? The fact not given which is needed to solve this problem is the

2-1 weather

2-2 date

2-3 time

2-4 miles covered

2-5 cost per gallon

Other items are (1) ability to read a graph, (2) size of fractions, (3) the facts needed to find the area of the front of a house, and (4) the conclusion that can be drawn from a bar diagram setting forth Federal expenditures for unemployment relief per year. Percentile norms are available for each part for grades 7, 8, and 9 based on the following:

| Grade | <i>N</i> |
|-------|----------|
| 7 | 1,564 |
| 8 | 2,241 |
| 9 | 3,773 |

The reliability of the total test and of its various parts was computed from 170 eighth-grade children. The narrowness of the range of one grade reduces the reliabilities somewhat. This test has also more value for predicting future success in algebra since the correlation between it and the Cooperative Elementary Algebra Test at the end of a year's study of algebra was reported as .78.

The following tests are useful for the survey of accomplishment in arithmetic: (1) the Compass Survey Tests of Arithmetic, and (2) Iowa Every-pupil Test of Basic Skills, Test D, Basic Arithmetic Skills.

DIAGNOSTIC TESTS IN ARITHMETIC

The following tests lay claim to being diagnostic tests in arithmetic: (1) the Compass Diagnostic Tests in Arithmetic, (2) the Diagnostic Test for Fundamental Processes in Arithmetic, and (3) the California Arithmetic Tests. Of these three, the Compass Diagnostic Tests in Arithmetic is by far the most comprehensive and complete in its coverage of arithmetic processes. It is probably the most efficient diagnostic test constructed in any subject. This test is divided into 20 different parts, as shown in the accompanying table. Each test has about five

¹ Item by permission of Educational Testing Service, Princeton, N.J.

COMPASS DIAGNOSTIC TESTS IN ARITHMETIC

| Grades | Time, minutes | Tests | Contents |
|--------|------------------|-------|---|
| 2-8 | 27 | I | Addition of whole numbers |
| 2-8 | 18 | II | Subtraction of whole numbers |
| 3-8 | 31 | III | Multiplication of whole numbers |
| 4-8 | 60 | IV | Division of whole numbers |
| 5-8 | 50 | V | Addition of mixed numbers |
| 5-8 | 40 | VI | Subtraction of mixed numbers |
| 5-8 | 30 | VII | Multiplication of mixed numbers |
| 5-8 | 40 | VIII | Division of mixed numbers |
| 5-8 | 45 | IX | Addition, multiplication, and subtraction of decimals |
| 6-8 | 40 | X | Division |
| 6-8 | 25 | XI | Addition and subtraction of denominate numbers |
| 6-8 | 30 | XII | Multiplication and division of denominate numbers |
| 7-8 | 54 | XIII | Mensuration |
| 6-8 | 38 | XIV | Basic facts of percentage |
| 7-8 | 44 | XV | Interest and business forms |
| 4-8 | 25 | XVI | Definitions, rules, and vocabulary of arithmetic |
| 5-6 | 35 | XVII | Problem analysis, elementary |
| 7-8 | 35 | XVIII | Problem analysis, advanced |
| 5-6 | 20 | XIX | General problem scale, elementary |
| 7-8 | 20 | XX | General problem scale, advanced |

parts. Some details about Test I will indicate the richness and completeness of the facts covered:

Part 1. 70 basic addition facts

Part 2. 66 higher decade addition facts

Part 3. 13 examples ranging from three to seven single digits of column addition

Part 4. 13 examples of more difficult column addition, from two two-place addition to seven three- to four-place numbers

Part 5. 7 examples similar to those in Part 4.

This test, so highly praised, has a few weaknesses. There is no indication of the reliability of the test as a whole or of that of any of its parts. There is perhaps an *inadequate treatment of arithmetic meanings*, for even the problems used are the traditional ones. Finally, it might be emphasized that such a diagnostic test locates the errors but does not arrive at the cause of the difficulty. It merely shows the level at which the pupil's work is unsatisfactory.

It is just at this point of understanding the cause of error that the Diagnostic Test for Fundamental Processes in Arithmetic by Buswell and John comes into the picture. This is an individual test in whose

administration a teacher sits down with a child and listens to him work aloud a carefully arranged set of examples. Its dominating purpose is to discover the reasons for the wrong habits. There are lists of types of errors which can easily be checked as the test proceeds. For example, in "addition" are listed:

1. Errors in combination
2. Counting
3. Added carried number last
4. Forgot to add carried number
5. Repeated work after partly done
6. Added carried number irregularly
7. Wrote number to be carried
8. Irregular procedure in column
9. Carried wrong number
10. Grouped two or more numbers

and eighteen other errors.

This test was the first to measure the thought patterns of children. However, it does not distinguish clearly between errors of computation and faulty work habits. The samples, too, are at times too few for real diagnosis. Some users have felt that the check list of errors is far from complete. This difficulty could be met by the teacher's writing down an account of the occasional error which did not occur in the list.

The third sample, the California Arithmetic Tests, already described under survey batteries, claims that it is a diagnostic test and offers procedures by which errors can be identified for the individual and summarized for the class as a whole. This test, a part of a battery, is divided into (1) reasoning and (2) fundamentals. When the test has been corrected and the scores brought forward to the first page a graph may be made of the scores, of the grade location, and of the percentile rank. If a child makes a poor record in arithmetic, his difficulties are then studied and a diagnostic analysis made of his learning difficulties. Both arithmetic reasoning and arithmetic fundamentals are analyzed into parts and each part keyed to the problem or example which illustrates it. For example, addition is analyzed into the following:

1. Sample combinations
 2. Bridging
 3. Carrying
 4. Zeros
 5. Column addition
 6. Adding money
 7. Adding numerators
- and five other parts.

It would be a distinct gain if a test could be used both as a survey and as a diagnostic test at the same time. There is no doubt, though, that the samples in this test are too few for a complete diagnosis. For example, there are only eight problems in long division scattered through three levels; errors in adding numerators are based on one example; in adding fractions and decimals, on one example, in denominated numbers, on one example; in adding fractions and decimals, on one example. The test is also weak in describing its manner of construction and perhaps in including such content as π , the square-root sign, and so on.

As a diagnostic instrument it suffers greatly in comparison with the Compass test. After describing the processes used in diagnosis one distinguished student of arithmetic¹ has written, "The uncritical user of tests should be protected from such spurious claims for 'diagnosis.'"

It might be said in extenuation of the claims of the tests that there are degrees of diagnosis. The study of the items of the usual survey test will give some explanation of the areas where little learning has taken place; the California Achievement Tests will add something more to the diagnosis, and the Compass Diagnostic Tests will go even farther than any of the above in getting at the root of the arithmetic difficulty. One must remember also that the hours of testing which are required in the Compass tests are far beyond the time that can usually be given to testing. However, the stimulation which an attempted analysis of errors gives is considerable, a fact which the author has recently learned from a study of the arithmetic errors made by pupils on the Metropolitan Achievement tests. For this reason he favors the effort made by the California Achievement Tests to analyze and classify as far as possible the errors of pupils.

TESTS OF MATHEMATICS IN HIGH SCHOOL

Tests suitable for testing the objectives of high school teaching of mathematics are described for the areas of algebra and geometry. Mention is also made of prognostic tests.

It is understood that the Cooperative Mathematics Test for Grades 7, 8, and 9 is also suitable for testing in the junior high school.

OBJECTIVES IN ALGEBRA TEACHING

The objectives in the teaching of algebra illustrate clearly two theories of the teaching of mathematics. The proponents of one theory emphasize the understanding of the symbols used in algebra, the learning of how to manipulate these symbols, the uses of the equation, and

¹ W. A. Brownell, *The 1938 Mental Measurements Yearbook* (Oscar K. Buros, ed.), Item 893. New Brunswick, N.J.: Rutgers University Press, 1938.

the understanding and use of graphs. The other group speak continuously of the process of generalization, of drawing inferences, and of exercising ingenuity in applying algebra to experiences which are now occurring or will be likely to occur.

A complete list of objectives must of necessity include the outcomes implied in the two theories just described. Breslich's list, for example, follows pretty closely the proponents of the first theory. According to him the major objectives are:¹

- a. To understand the terminology of the algebra taught during the semester.
- b. To perform the fundamental operations that have been taught.
- c. To combine and decompose simple algebraic expressions.
- d. To derive equations from problems.
- e. To solve equations.
- f. To understand formulas.
- g. To evaluate formulas.
- h. To solve formulas for a given letter.
- i. To translate verbal statements into formulas.
- j. To understand graphical representation.
- k. To use graphical representation.

It is quite evident that these objectives should be supplemented by the following:

1. The ability to draw inferences from algebraic data
2. The capacity to generalize from mathematical facts presented
3. Ingenuity in applying mathematical techniques to practical problems
4. The ability to synthesize and coordinate mathematical facts through the process of thinking
5. The capacity to select data and bring it to bear on the solution of the problem at hand

It is indeed a difficult problem to construct tests which measure adequately these latter objectives. If a test is suitable only for checking the first set of objectives, then this fact should appear in the title. Perhaps instead of describing a test as "an algebra test," it should be described as "a test of the manipulative and mechanical aspects of algebra." This description might appear in the subtitle. At any rate, some descriptive statement should appear to inform the user that the testing of the whole area of algebra instruction was not contemplated.

¹ See Buros, Oscar K. (ed.), *The Nineteen Forty Mental Measurements Yearbook*, Item 1435. Highland Park, N.J.: The Mental Measurements Yearbook, 1941. By permission.

If these recommendations were carried out, nine-tenths of the criticisms of these instruments would not be necessary.

ALGEBRA TESTS

Two algebra tests will be discussed: (1) the Columbia Research Bureau Algebra Tests, and (2) the Cooperative Algebra Test.

The Columbia Research Bureau Algebra Test for grades 9 and 13 is divided into two levels:

| Test | Time, minutes |
|------------------------------|---------------|
| I. First semester..... | 80 |
| II. Revised, first year..... | 100 |

Test I is divided into Part I, Mechanics, and Part II, Problems. Part I consists of 17 equations and 3 graph exercises to be solved. Most of them are short and simple:¹

5. $\frac{x}{5} = 1.2$

11. $6x^2 - 13x + 6 = 0$

15. $F = \frac{9x}{5} + 32$

19 = Plot in Chart A the graph of the equation $x + 2y = .4$

The outline graph for Item 19 is furnished. Part I includes also factoring, treatment of signs, and a factorable quadratic equation.

In Part II, the student must write the correct equation for each of 20 problems and then solve it correctly.

2. If a piece of cloth 44 inches long will shrink to 42 inches when washed, how many inches long will a 33 inch piece of the same cloth be after shrinking?
15. How many dollars put at simple (*i.e.*, not compound) interest for 2 years at $5\frac{1}{2}\%$ per annum will amount to \$100?

On the next to the last page there are lines on which are to be entered the equations of the problems as well as the value of the unknowns.

Test II is much like Test I in form but the equations, exercises, and problems are longer and more complicated. The two forms of Test I correlate with each other .89 to .94. The reliability is given at .94. Norms are based on moderately small numbers. For example, the norms of Test I were based on the records of 598 students who had just finished a semester of algebra. The feature of the test whereby the shorter more mechanical items of Test I are supplemented by the longer more complicated exercises and problems in Test II is an excellent one. The test reflects the teaching of objectives of a more

¹ Items by permission of World Book Company, Yonkers, N.Y.

mechanical and manipulative nature. Correlations ranging from .68 to .72 between the test scores and teachers' marks are reported.

The Cooperative Algebra Test was carefully constructed and later revised.¹ At present there are many forms. Each of which is divided into three parts. Two levels of difficulty are provided for. One test, Elementary Algebra through Quadratics, is the simpler one and Quadratics and Beyond, the more complex one. The form of the items of both tests is that of multiple choice. The examples and problems differ slightly from those ordinarily experienced in the usual algebra in both form and lettering. The test, Elementary Algebra through Quadratics is divided into three parts. Part I contains 20 samples of algebraic manipulations. Items containing the collection of terms, uses of negative numbers, removal of parentheses, solution of equations, simplifying fractional terms, and treatment of exponents in multiplication appear. What the subject is to do is made clear in each item. Here are two items from Form Q which illustrate both the materials of this part and the technique used:

5. The sum of $-15c^4$ and $-3c^4$ is

5-1 $-18c^4$

5-2 $-12c^4$

5-3 12

5-4 $12c^8$

5-5 $-18c^8$

19. If the graph of the equation $3x + 5y = 1$ passes through the point $(m, -4)$ the value of m is

19-1 $2\frac{3}{5}$

19-2 -7

19-3 $6\frac{1}{3}$

19-4 7

19-5 $-2\frac{1}{5}$

Part II deals with the solution of 15 problems, two of which are graphs. The solution of problems involving percentage, constructing equations, and ratio is called for. The type of problems used is shown by two samples.

4. In a certain high school there are 200 more girls than boys. The total number of pupils in the school is 1876. How many boys are there?

4-1 638

4-2 838

4-3 1038

4-4 1138

4-5 1676

¹ Items by permission of Educational Testing Service, Princeton, N.J.

15. A dealer wishes to mix hazelnuts worth 50¢ a pound and cashews worth 75¢ a pound to obtain 10 lbs. of mixed nuts worth 55¢ a pound. How many pounds of cashews would he use?

15-1 $3\frac{1}{2}$

15-2 2

15-3 $6\frac{2}{3}$

15-4 4.2

15-5 8

Part III contains 28 items which involve algebraic manipulation of formulas and equations which, for the most part, involve symbols instead of numbers. Two samples are:

9. If $\frac{s}{a} = \frac{1}{h}$, then s equals

9-1 $\frac{1}{h} - \frac{1}{a}$

9-2 $\frac{a}{h}$

9-3 ah

9-4 $\frac{1}{ah}$

9-5 $\frac{h}{a}$

14. If $Sx - 7 = cx$, then x equals

14-1 $S - \frac{c}{7}$

14-2 $c - \frac{7}{S}$

14-3 $\frac{7}{S - c}$

14-4 $\frac{-7}{Sc}$

14-5 $\frac{S - c}{-7}$

Two fundamental criticisms have been leveled at the Cooperative test. The first one claims that the test is too mechanical, that it fails because it emphasizes too much the mechanics and manipulative aspects of algebra. The test is weak in its measurements of the ability to draw inferences from data and of the ingenuity needed in applying algebraic techniques to practical problems. Thus the ability to synthesize and coordinate mathematical facts through the process of thinking is largely neglected. In answer to such castigations the Cooperative Test Service has said that there are certain facts which the test constructor must consider. These might be formulated in a set of questions:

1. Does a large majority of authorities agree as to the importance of the objectives?
2. Are the teachers actually striving to achieve these objectives?
3. Is the objective clear, specific, and unambiguous?
4. Is the objective capable of immediate attainment?

In regard to the Cooperative Algebra Test it might be said that the authorities agree generally on the importance of the items. From an inspection of the algebra used, there is no doubt that the teachers are striving to achieve them. The clarity and specificity of the problems are beyond question and are immediately attainable.

It may be concluded, therefore, that the test measures well what it sets out to measure and hence is valid for that purpose. It does not emphasize those higher outcomes of thinking, inference, and application. Perhaps instead of calling it the Cooperative Algebra Test it could be more properly called "the Cooperative Algebra Test of the Mechanical and Manipulative Aspects of Algebra."

The second criticism, possibly not so significant, is aimed at the form in which the items of this test are cast—the multiple-choice form. Mathematicians object to the guessing involved. They say that it does not test computational accuracy or the capacity to select and synthesize data or to coordinate thought. The authors of the Cooperative Algebra Test realized this difficulty. They actually introduced answers which would mislead superficial inspection. These answers appear plausible to those who are deficient in the very ability which is being tested. It cannot be denied that this objection is to a certain extent valid, but it must be balanced against the ease of scoring which is inherent in the multiple-choice form.

GEOMETRY TESTS

The same differences of emphasis characterize the objectives of geometry as was the case with algebra. One group of students would delete large areas of the usual textbook material in plane geometry. Its members would keep only those theorems which have practical meaning for those students who will not use geometry technically. They emphasize the nature of proof. Applications of the rigor of proof exemplified in geometry they would have applied to the problems of the day until the student knew what good argument is like. Transfer of training in geometry under such a regime of instruction could be expected to take place.¹ The other group believes that the theorems and problems of the usual textbook are, after all, what can be learned. They would have this as full of meaning as possible, but they would not de-

¹ See review by Leroy N. Schnell in *The Nineteen Forty Mental Measurements Yearbook op. cit.*, Item 1467.

lete great areas. This latter group would agree pretty largely with the objectives developed by Lide in his book, *Instruction in Mathematics*:¹

1. Development of logical reasoning ability.
2. Development of an appreciation of the utility and beauty of geometrical forms.
3. Familiarization of the student with the properties, mensuration, and relationships of common geometric forms.
4. Development of an understanding and an appreciation of deductive proofs.
5. Creation of an understanding of spatial concepts and relations.
6. Establishment of habits of precision and accuracy.
7. Development of an appreciation of the part geometry has played in the history of civilization.

Only a few tests are able to measure the rich variety of objectives set forth above.

Achievement Tests in Plane Geometry

Perhaps the *Examination in Plane Geometry*, high school level,² measures as many of the objectives in plane geometry as any other test. This test of geometry is divided into three sections:

Section I. Ability to use proofs and to arrive at logical conclusions.

Section II. Ability to handle various types of simple constructions.

Section III. Ability to solve practical, everyday problems involving geometric principles and facts.

This test, developed for use in the United States Army, is suitable for the tenth grade and consumes 1 hour and 40 minutes of working time. The percentile norms for the total score are available.

Much less comprehensive is the Cooperative Plane Geometry Test which takes 40 minutes to administer.³ This test also has three parts.

Part I contains 30 items to be marked "true" or "false" and consumes 10 minutes. It is essentially a test of geometric information.

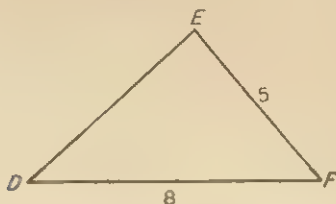
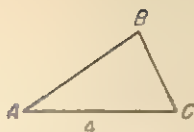
Part II consists of 20 theorems, or problems, which deal with the circle, equality and similarity of triangles, the rhombus, and the hexagon. Fifteen minutes are used in administration. Sample problems are from Form Q:

¹ Lide, Edwin S., *Instruction in Mathematics*, National Survey of Secondary Education, Monograph No. 23 (U.S. Office of Education, Bulletin 1932, No. 17). Washington, D.C.: Government Printing Office, 1933.

² Educational Testing Service, Princeton, N.J.

³ Items by permission of Educational Testing Service, Princeton, N.J.

5.



If triangle ABC is similar to triangle DEF, and if $AC = 4$, $DF = 8$ and $EF = 5$, then BC equals

5-1 1

5-2 $2\frac{1}{2}$ 5-3 $6\frac{2}{3}$

5-4 10

5-5 Solution impossible

20. Two parallel chords of a circle are each 16 inches in length; the distance between them is 12 inches. The radius of the circle is

20-1 10 inches

20-2 20 inches

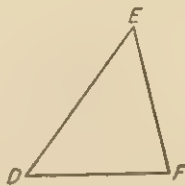
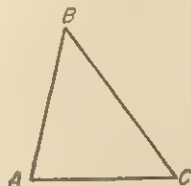
20-3 6 inches

20-4 8 inches

20-5 5 inches

Part III, which also requires 15 minutes for taking, consists of 15 problems more complicated than those in Part II. They deal with circles, triangles, and parallelograms.

5.



Given angle A = angle D

$$\frac{AB}{DE} = \frac{AC}{DF}$$

Angle C can be proved equal to Angle F by proving from the given facts that the two triangles are

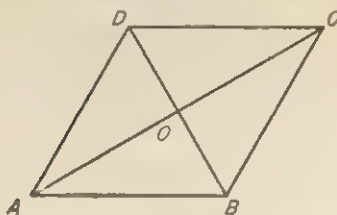
5-1 similar

5-2 congruent

5-3 equiangular

5-4 equilateral

5-5 equal in area



12.

Given: Parallelogram ABCD with diagonals AC and BD intersecting at O.
 One of the simplest proofs that $AO = OC$ uses the statement that

12-1 alternate interior angles of parallel lines cut by a transversal are equal.

12-2 corresponding angles of parallel lines cut by a transversal are equal.

12-3 the opposite angles of a parallelogram are equal.

12-4 adjacent sides of a parallelogram are equal.

12-5 if two lines are parallel, the interior angles on the same side of the transversal are supplementary.

Scaled scores are provided and percentile norms for high school classes in plane geometry are available.

While this test emphasizes very little the nature of proof or the applications of geometry to life situations it does test well the specific attainable content of the more usual type of geometry instruction. It seems to the author that the true-false form used in this test is a distinct limitation. Furthermore, the multiple-choice form does not permit the student to marshal his own arguments and select that one which is most fitting. One must not forget, however, that multiple-choice answers can be machine-scored and in this manner inordinate amounts of time saved.

Many of the transfer values of geometry have been neither agreed upon nor clearly defined. Until they are both, measurement in this area will be hindered. Some mathematicians object strenuously to limiting the proof to one procedure. They believe that fixing the number of steps in the proof stifles ingenuity.

Because the nature of proof looms so large in the teaching of geometry there is a distinct impression gained from reading the criticisms of mathematicians that standardized tests in geometry are not so satisfactory as those in algebra. It seems impossible at the present time to allow the subject sufficient latitude of choice in the formalized proof of the standard test.

PROGNOSTIC TESTS IN ALGEBRA AND GEOMETRY

When pupils are slow in acquiring arithmetic and low in their intelligence-test scores, should they continue their mathematical training in the more formal courses of algebra and geometry? The answer to this question depends on the pupil's interest, his vocational plans, and the

length of time he intends to stay in school. In helping the child make such a decision the counselor needs all the information which can be collected. To help answer such a question there have been developed prognostic tests which help to foretell a child's standing in algebra or geometry. Because these tests prophesy only in a moderate degree anticipated scores or marks, they must of necessity be used only as added information. If the prognostic test, achievement tests in arithmetic, and intelligence test agree with school marks, then the prognosis is less likely to be incorrect.

Prognostic Tests in Algebra

Two types of prognostic tests have been constructed. One of these is more dependent upon what the pupil has learned; the other, on his capacity to learn material similar to what the course actually contains. The Lee Test of Algebraic Ability illustrates the first; the Orleans Algebra Prognosis, the second. Let us look more closely at the latter. This test by Orleans is divided into a test on arithmetic and into 12 other parts: (1) substitution in monomials, (2) use of exponents, (3) meaning of exponents, (4) substitution in monomials with exponents, (5) substitution in binomials with exponents, (6) like and unlike terms, (7) representation of relations, (8) representation of expressions, (9) positive and negative numbers, (10) problems, (11) addition of like terms, and (12) summary test. Each part except No. 12 contains both a lesson and a test thereon. Let us look at Lesson 3. There are seven illustrations of how to deal with exponents. Item 3 reads, " a^2 means a times a . If $a = 3$, then a^2 means 3^2 or 3×3 , which equals 9." In the test you are advised that you may look back to Lesson 3 if you need to. Item 4 of Test 3 is "What does c^3 mean?" Lesson 7 has 9 items on how to use positive and negative numbers. Item 4 in this lesson says, " -12 followed by $+3$ means a loss of 12 followed by a gain of 3, which results in a net loss of 9. This is written $-12 + 3 = -9$." Item 4 in Test 9 is " $-10 - 2$." The fundamental question is how much algebra a student can learn in a defined amount of time (81 minutes).

The validity of this test has been determined by measuring the prophecy as obtained from the prognosis test against achievement $4\frac{1}{2}$ months later as measured by a standardized achievement test of algebra. In one case this correlation was .82; in another, .71. Since the test is now appreciably longer than when these computations were made, the authors believe the coefficient to be about .80 on the average. Do not forget that a correlation of .80 means an efficiency only 40 per cent better than chance.

In geometry, similar prognostic tests have been constructed by the Orleans, by Lee, and by the Iowa authors.

SUMMARY

Objectives in the teaching of mathematics vary from the learning of the meaning and manipulation of mathematical symbols to the learning of mathematical reasoning as applied to the problems of daily life.

In the elementary school these objectives become those of teaching pupils the fundamentals of quantitative thinking. Pupils must learn to perform quickly and well the four fundamental operations with integers and fractions, to have some understanding of quantity in business transactions, and to understand the quantitative aspects of problems arising in daily life. In the elementary school this means learning the fundamentals of arithmetic.

Tests of these arithmetic processes are furnished in the survey batteries which test also the other areas of instruction. These commercially available standardized tests usually have sections both on arithmetic fundamentals and on arithmetic problems. In most cases there are opportunities for some analysis of each individual's strong and weak points. These batteries usually have arithmetic tests available at all levels of progress. For more complete testing of progress in arithmetic there are the separate batteries. There are also excellent diagnostic tests of arithmetic, both group and individual. All told, progress toward defined objectives is well measured in the area of arithmetic.

At the high school level there are available tests of general mathematics, algebra, geometry (both plane and solid), and trigonometry. Of these, the tests of algebra seem most valid. In the area of algebra a test which satisfactorily measured the acquisition of the four fundamental operations, the applications of formulas, and the formulation and solution of the equation was severely criticized because it neglected to test the capacity to generalize or the ability to synthesize and coordinate mathematical facts. For this reason it was suggested that more refined descriptions of the nature of the tests be furnished the users. Likewise in geometry the tests failed to emphasize the nature of proof or the applications of geometry to life situations. These processes are placed high among desirable objectives by modern teachers.

LIST OF MATHEMATICAL TESTS

I. ARITHMETIC TESTS

Survey

1. The Cooperative Mathematics Test for Grades 7, 8, and 9. 1940. Several forms. Time: 80 minutes. Reliability: .92. Authors. Alice H. Darnell, John C. Flanagan, Stevenson W. Fletcher, and

Rose E. Lutz. Cooperative Test Service, New York.

2. Arithmetic tests in the following batteries: (a) Stanford Achievement Tests, (b) Metropolitan Achievement Tests, (c) California Achievement Tests, and (d) Coordinated Scales of Attainment Tests.

3. Analytical Scales of Attainment in Arithmetic, grades 3-4, 5-6, 7-8. 1933. Two forms. Three levels. Time: 80 minutes. Authors: L. J. Brueckner, Martha Kellogg, and M. J. Van Wagenen. Educational Test Bureau, Minneapolis, Minn.

4. Compass Survey Tests in Arithmetic, grades 2-8. 1927. Two forms. Two levels. Grades 2-4, 25 minutes; grades 4-8, 35 minutes. Authors: H. A. Greene, F. B. Knight, G. M. Ruch, and J. W. Studebaker. Scott, Foresman & Company, Chicago.

5. Basic Arithmetic Skills, Iowa Every-pupil Tests of Basic Skills. New edition, 1940, 1945. Forms L, M, N, and O. Two levels. Elementary battery, grades 3-5, 57 (65) minutes. Advanced battery, grades 5-9, 68 (80) minutes. Form O machine scored. Author: N. F. Spitzer aided by Ernest Horn, H. A. Greene, and E. F. Lindquist. Houghton Mifflin Company, Boston.

Diagnostic Tests

1. Compass Diagnostic Tests in Arithmetic, grades 2-8. 1925. One form. 20 parts. Time for each part ranges from 18 to 54 minutes. Authors: G. M. Ruch, F. B. Knight, H. A. Greene, and J. W. Studebaker. Scott, Foresman & Company, Chicago.

2. Diagnostic Test for Fundamental Processes in Arithmetic, grades 2-8. 1925. An individual test. Two forms. Nontimed (about 20 minutes). Authors: G. T. Buswell and Lenore John. Public School Publishing Company, Bloomington, Ill.

3. California Arithmetic Tests. 1933-1939. Two forms. Three levels. Primary battery, grades 2-3, 50 minutes; elementary battery, grades 4-6, 60 minutes; intermediate battery, grades 7-9, 75 minutes; advanced battery, grades 9-14, 68 minutes. Authors: Ernest W. Tiegs and Willis W. Clark. California Test Bureau, Los Angeles, California.

4. Diagnostic Tests in Arithmetic Fundamentals, grades 2-6. 1945. One

form. Five levels. Different material for each grade. grade 2, addition and subtraction, 87 (110) minutes; grade 3, addition, subtraction, and multiplication, 73-95 minutes; grade 4, Part 1, addition and subtraction, 100 (120) minutes; grade 4, Part 2, multiplication and division, 90 (110) minutes; grade 5, Part 1, addition, subtraction, multiplication, and division, 80 (100) minutes; grade 5, Part 2, fractions (addition and subtraction), 90 (110) minutes; grade 6, Part 1, addition, subtraction, multiplication, division, 60 (75) minutes; grade 6, Part 2, fractions and decimals, 75 (95) minutes. Authors: Department of Educational Research, Ontario College of Education, University of Toronto.

5. Hundred Problem Arithmetic Test, grades 7-12. 1926-1944. Forms V and W. Time: 40 (45) minutes. Authors: Raleigh Schorling, John R. Clark, and Mary A. Potter. World Book Company, Yonkers, N.Y.

II. ALGEBRA TESTS

1. Breslich Algebra Survey Test, high school. 1930-1931. First semester, 41 minutes; second semester, 52 minutes. Author: E. R. Breslich. Public School Publishing Company, Bloomington, Ill.

2. Columbia Research Bureau Algebra Test, grades 9 or 13. 1927-1933. Two forms. Two levels. Test 1, first semester, grade 9 or 13. 80 minutes. Test 2, revised, first year, grades 9-14, 100 minutes. Authors: Arthur S. Otis and Ben D. Wood. World Book Company, Yonkers, N.Y.

3. Snader General Mathematics Test, grade 9. 1951. Two forms, AM and BM. Time: 40 minutes. Reliability: .80 and .84. Norms for end of year based on 2,190 students in 22 states, C.A. 15-4, I.Q. 98. Arithmetic 42 per cent; informal geometry, 23 per cent; graphic representation, 8 per cent; algebra, 25 per cent; numerical trigonometry, 2 per cent. Evaluation and Adjustment Series edited by Walter N. Durost. World Book Company, Yonkers, N.Y.

4. Cooperative Algebra Test, Elementary Algebra through Quadratics, revised series, high school. 1937-1943. Forms Q, R, S, and T. Machine scored but separate answer sheets need not be used. Time: 40 (45) minutes. Scaled scores are provided. Authors: John A. Long, L. P. Siceloff, Leone E. Cheshire, Margaret P. Martin, and Marion F. Shaycoft. Cooperative Test Service, New York.

5. Cooperative Intermediate Algebra Test, Quadratics and Beyond, revised series, high school. 1941-1943. Forms R, S, and T. Time: 40 minutes. Authors: John A. Long, L. P. Siceloff, Leone E. Cheshire, and Marion F. Shaycoft. Cooperative Test Service, New York.

6. Lankton First Year Algebra Test, high school. 1951. Forms AM and BM. Time: 40 minutes. Reliability: .84 and .87. Percentile norms based on 3,183 students from 22 states. Median C.A. of students 15-1, median I.Q. 106. Simple operations, formulas, equations, graphs, problem solving. Evaluation and Adjustment Series edited by Walter N. Durost. World Book Company, Yonkers, N.Y.

7. Iowa Every-pupil Test in Ninth Year Algebra, high school. New form each May. Time: 55 minutes. Author: H. Vernon Price. Bureau of Educational Research and Service, State University of Iowa, Iowa City.

III. ACHIEVEMENT TESTS IN PLANE AND SOLID GEOMETRY AND TRIGONOMETRY

Plane Geometry

1. Cooperative Plane Geometry Test, high school. 1933-1940. Forms at present O, P, Q, R, S, and T. Time: 40 minutes. Three parts. Scaled scores are provided. Percentile norms are available. Authors: Emma Spaney, L. P. Siceloff, *et al.* Cooperative Test Service, New York.

2. Davis Test of Functional Competence in Mathematics, grades 9-12. 1951. Two forms, AM and BM. Time: 40 minutes. Reliability: .81 to .91.

Grade 9, C.A. 15-2, I.Q. 100; grade 10, C.A. 16, I.Q. 102; grade 11, C.A. 17, I.Q. 103; grade 12, C.A. 17-11, I.Q. 105. Consumer problems, problems of rent, insurance, changing money, investment, bonds, banking, budgeting, etc. Evaluation and Adjustment Series edited by Walter N. Durost. World Book Company, Yonkers, N.Y.

3. Orleans Plane Geometry Achievement Test. 1929. Two equivalent forms. Test 1, for first semester, covers Books I and II except loci; Test 2, for second semester, covers Books III, IV, and V and loci. Percentile norms based on 3,500 cases are available. Reliability: Test 1, .85; Test 2, .71. Authors: Joseph B. and J. S. Orleans. World Book Company, Yonkers, N.Y.

4. Shaycoft Plane Geometry Test, high school. 1951. Forms AM and BM. Time: 40 minutes. Reliability: .82. Percentile norms based on 2,914 students in 24 states. Median C.A. 16-2, I.Q. 110. Adds analytic versus synthetic proofs and indirect proof (2 per cent). Small attempt at geometric reasoning. Evaluation and Adjustment Series edited by Walter N. Durost. World Book Company, Yonkers, N.Y.

5. Columbia Research Bureau Plane Geometry Test. 1926. Two equivalent forms. Working time: 60 minutes. Reliability between two forms: .93. Authors: Herbert E. Hawkes and Ben D. Wood. World Book Company, Yonkers, N.Y.

6. Iowa Plane Geometry Aptitude Test, revised edition, high school. 1935-1942. One form. Time: 44 (50) minutes. Machine-scored, though separate answer sheets need not be used. Authors: H. A. Greene and H. W. Brace. Bureau of Educational Research and Service, State University of Iowa, Iowa City.

Solid Geometry

1. Cooperative Solid Geometry Test, high school. 1932-1938. Forms O and P. Scaled norms. Percentile norms for high school classes in solid geometry. Time: 40 minutes. Authors: H. T. Lundholm,

John A. Long, and L. P. Siceloff. Cooperative Test Service, New York.

Trigonometry

1. Cooperative Trigonometry Test, revised, grades 11–15. 1928–1930. Forms O, P, and U. Time: 40 minutes. Scaled and percentile scores provided for high school and college classes in trigonometry. Authors: John A. Long and L. P. Siceloff. Cooperative Test Service, New York.

IV. PROGNOSTIC TESTS IN GEOMETRY

1. Iowa Plane Geometry Aptitude Test, high school. 1935. One form. Time: 44 minutes. Correlation between aptitude test and a 90-minute objective achievement test: .70. Correlation with the average of first-semester and second-semester school marks combined: .59. Authors: Harry A. Greene and Harold W. Bruce. Bureau of Educational Research and Service, State University of Iowa, Iowa City.

2. Lee Test of Geometric Aptitude, high school. 1931. One form. Time: 31 minutes. Median correlation between this test of geometric aptitude and achievement-test score: .765. Correlation between aptitude test and school marks: .53. Reliability: .81 (N , 107). Authors: Doris M. Lee and A. Murray Lee. California Test Bureau, Los Angeles, Calif.

3. Orleans Geometry Prognosis Test, high school. 1929. One form. Time: 70 minutes. Correlation between the prognostic battery and an achievement test: .73 (probably would be raised to .80 with the present much-lengthened test). No reliability reported. Authors: Joseph B. and Jacob S. Orleans. World Book Company, Yonkers, N Y.

V. PROGNOSTIC TESTS OF ALGEBRA

1. Iowa Algebra Aptitude Test, grade 9. 1931. One form. Correlation with single achievement test: .66 (N = 105). Probably more information needed about its construction and validation. Time: 35 minutes. Authors: Harry A. Greene and Alva H. Piper. Bureau of Educational Research and Service, State University of Iowa, Iowa City.

2. Lee Test of Algebra Ability, grade 9. 1930. One form. Time: 25 minutes. Correlation between this test and test of achievement: .71. Reliability: .93 (split-half method). Author: J. Murray Lee. Public School Publishing Company, Bloomington, Ill.

3. Orleans Algebra Prognosis Test, grades 7–9. 1928–1932. One form. Time: 81 minutes. Correlation with achievement test at end of semester: .71 and .82 (.80 estimated with present length). No reliability reported. Authors: Joseph B. and Jacob S. Orleans. World Book Company, Bloomington, Ill.

QUESTIONS AND EXERCISES

1. *a.* Describe the objectives in teaching arithmetic.

b. How far do you think such objectives are measured by the tests of fundamentals and problems contained in the general test batteries?

2. Secure a copy of the California Achievement Tests and study their provisions for diagnosis. Do you think such an instrument is adequate for the purpose of diagnosis?

3. Describe in some detail the Compass Diagnostic Tests in Arithmetic. Make a detailed study of this instru-

ment and conclude as to whether the author of this test was justified in calling it "the most efficient diagnostic test constructed in any subject."

4. What are the leading characteristics of the Buswell-John Diagnostic Test for Fundamental Processes in Arithmetic? The Cooperative Mathematics Test for Grades 7, 8, and 9?

5. Why should a few students not take algebra or geometry?

6. Compare in some detail the two points of view present in formulating objectives in algebra and geometry.

7. Summarize the leading characteristics of the Cooperative Mathematics Test for Grades 7, 8, and 9.

8. What are the characteristics of an excellent diagnostic test in arithmetic? What are the limitations of diagnosis in a survey test?

9. Describe and illustrate the two types of objectives in the teaching of algebra.

10. What are the characteristics of the Cooperative Algebra Test? Describe the criticisms leveled at this test and evaluate them.

11. What are the major outcomes of the teaching of geometry?

12. Evaluate the criticisms of the multiple-choice technique in constructing geometry tests. What outcomes in the teaching of geometry does this technique fail to measure?

13. Discuss the uses of prognostic tests in guidance. How valuable are the tests?

14. What are the two types of prognostic tests of mathematics? Describe one of them.

BIBLIOGRAPHY

Books

BUROS, OSCAR K. (ed.): *The Nineteen Forty Mental Measurements Yearbook*, Items 1431-1475. Highland Park, N.J.: The Mental Measurements Yearbook, 1947.

———: *The Third Mental Measurements Yearbook*, Items 303-362. New Brunswick, N.J.: Rutgers University Press, 1949.

BUSWELL, G. T., with the cooperation of LENORE JOHN: *Diagnostic Studies in Arithmetic*, Supplementary Educational Monographs No. 30, University of Chicago, 1926.

COMMISSION ON SECONDARY SCHOOL CURRICULUM, PROGRESSIVE EDUCATION ASSOCIATION: *Mathematics in General Education*, Chap. XIII. New York: Appleton-Century-Crofts, Inc., 1940.

The Cooperative Achievement Tests—A Handbook—1936. New York: Cooperative Test Service.

GREENE, H. A., A. N. JORGENSEN, and J. R. GERBERICH: *Measurement and Evaluation in the Secondary School*, Chap. XVIII. New York: Longmans, Green & Co., Inc., 1943.

HAWKES, H. E., E. F. LINDQUIST, and C. R. MANN (eds.): *The Construction and Use of Achievement Tests*, Chap. VII. Boston: Houghton Mifflin Company, 1936.

LIDE, EDWIN S.: *Instruction in Mathematics*, National Survey of Secondary

Education, Monograph No. 23 (U.S. Office of Education, Bulletin 1932, No. 17). Washington, D.C.: Government Printing Office, 1933.

LINDQUIST, E. F. (ed.): *Educational Measurement*, Chaps. 1, 2. Washington, D.C.: American Council on Education, 1951.

O DELL, C. W.: *Educational Measurements in High School*, Chap. VII. New York: Appleton-Century-Crofts, Inc., 1930.

SYMONDS, P. M.: *Measurement in Secondary Education*, Chap. VI. New York: The Macmillan Company, 1928.

Articles

"Arithmetic in General Education," *Sixteenth Yearbook of the National Council of Teachers of Mathematics*. New York: Bureau of Publications, Teachers College, Columbia University, 1941.

BECKER, IDA S.: *The Construction and Standardization of a Test in Plane Geometry*, unpublished master's thesis, Kansas State Teachers College, 1934.

COOKE, DENNIS H., and JOHN M. PEARSON: "Predicting Achievement in Plane Geometry," *School Science and Mathematics* (1933) 33:872-878.

GROVER, C. C.: "Results of an Experiment in Predicting Success in First Year Algebra in Two Oakland Junior High Schools," *Journal of Educational Psychology* (1932) 23:309-314.

LEE, J. MURRAY, and DORIS MAY LEE: "The Construction and Validation of a Test of Geometric Aptitude," *Mathematics Teacher* (1932) 25:193-203.

ORLEANS, JOSEPH B.: "A Study of Prognosis of Probable Success in Algebra and in Geometry," *Mathematics Teacher* (1934) 27:165-180, 225-246.

——— and P. M. SYMONDS: "The Comparative Reliabilities of Standardized and Teacher-made Achievement Tests When Given in the Middle of the Year," *Journal of Educational Research* (1933) 25:127-128.

PERRY, WINONA M.: "Prognosis of Abilities to Solve Exercises in Geome-

try," *Journal of Educational Psychology* (1931) 22:604-609.

PIPER, A. H.: *The Validity of Certain General and Special Tests for Prognosis in First Year Algebra*, unpublished master's thesis, State University of Iowa, 1929.

SEAGOE, MAY V.: "Prediction of Achievement in Elementary Algebra," *Journal of Applied Psychology* (1938) 22: 493-503.

TORGERSON, T. L., and GENEVA P. AAMODT: "The Validity of Certain Prognostic Tests in Predicting Algebraic Ability," *Journal of Experimental Education* (1933) 1:277-279.

CHAPTER 10

Measurement of Science

Science and scientific thinking have come to form such an integral part of daily life that their understanding becomes one of the major objectives of education. Probably in no other areas of learning are there more opportunities for application and illustration. The very problems of living and breathing, of health and recreation, of buying and selling, of transportation and social interaction are scientific problems so profuse that great difficulty has been experienced in agreeing on a unified course of study.

Just as in the social sciences so in the natural sciences there are the problems of learning meaningful facts and their translation into life patterns. The application of the scientific method to the solution of everyday problems thus becomes one of the most important outcomes of the educational process.

AIMS AND OBJECTIVES OF SCIENCE TEACHING

The objectives of instruction in science are divided into two parts: (1) the learning and understanding of scientific facts; (2) the development of the scientific method.¹

1. Learning and understanding of facts and information acquired in the various sciences
 - a. The discovery of illustrations in daily life
 - b. Explaining and understanding of ordinary problems in daily life, which frequently involve the application of generalizations learned in the classroom
 - c. Making predictions about the outcomes of problems based on the learned facts and principles
 - d. Ability to read and understand scientific materials
 - e. Mastery of the terms and concepts peculiar to work in science
 - f. Skill in laboratory techniques

¹ These objectives parallel very closely but not exactly those set forth in "The Measurement of Understanding in Science," Chap. VI, *Forty-fifth Yearbook of the National Society for the Study of Education*, Part I, "The Measurement of Understanding." Chicago: University of Chicago Press, 1946.

- g. Familiarity with well-authenticated sources of information
- h. Ability to name forms or structures and processes and to be acquainted with their functions.
- 2. Developing the scientific method
 - a. Making the proper qualifications when interpreting data
 - (1) Staying within the limits of the facts presented
 - (2) Using caution and reservation in the inferences drawn
 - (3) Avoiding the influence of irrelevant facts
 - b. Ability to interpret data, *i.e.*, to recognize trends in data by seeing common elements in diverse data
 - c. Ability to identify valid cause-and-effect relationships
 - d. Ability to draw correct conclusions from scientific data
 - e. Giving correct reasons which adequately support conclusions
 - (1) Knowing and selecting the principle that applies to the situation
 - (2) Avoiding the influence of irrelevant factors
 - (3) Citing reliable authorities
 - (4) Avoiding both popular misconceptions and the assumption of conclusions
 - f. Ability to formulate hypotheses and to plan experiments to test them
 - g. Ability to identify the assumptions, whether stated or not, which are necessary to draw the conclusion.
- 3. To develop in children habits of healthful living, which include habits of performing useful tasks and of applying scientific principles in daily life
- 4. To develop in children interest in the scientific problems around them and in science itself
- 5. To develop in children some appreciation of the beauties of nature and of commonplace events which are so easily taken for granted.

As we examine the tests we shall raise the question as to what aspects of their teaching aims and objectives are measured by the instrument in question. We shall first examine tests suitable for the testing of the objectives of science teaching in the elementary school and second, in the high school.

TESTS OF SCIENCE IN THE ELEMENTARY SCHOOL

Tests of science appear in several of the achievement test batteries suitable for testing the outcomes of instruction in the elementary school. However, in such batteries as the California Achievement Test and the Iowa Every-pupil Tests of Basic Skills, which concentrate on the testing of basic skills, there are no tests of science achievement.

SCIENCE TESTS IN TEST BATTERIES

The science tests occurring in three batteries will be described: (1) the Coordinated Scales of Attainment, (2) the Stanford Achievement Test, and (3) the Metropolitan Achievement Tests.

The Coordinated Scales of Attainment are so constructed that there is a separate test battery for each grade.¹ For this reason, opportunity is given for a larger and more complete coverage of the area of science information than is true of any other battery. Let us look at the tests used in grades 4, 5, and 6. In most test batteries there would be only one test for all three grades, which would usually contain 50 to 60 items. In the Coordinated Scales of Attainment, however, there are 60 items for each grade or 180 items for the three. If we group grades 4, 5, and 6 into one division and grades 7 and 8 in another, the contents of each group may be very roughly classified as shown in Table 8. This table

TABLE 8. CONTENTS OF SCIENCE TESTS

| Subject | Average number of items | | | | | |
|------------------------|----------------------------------|------------|---------------------------|------------|-------------------------------|------------|
| | Coordinated Scales of Attainment | | Stanford Achievement Test | | Metropolitan Achievement Test | |
| | Grades 4-6 | Grades 7-8 | Grades 4-6 | Grades 7-8 | Grades 4-6 | Grades 7-9 |
| Animals..... | 13 | 5 | 12 | 11 | 13 | 9 |
| Plants..... | 10 | 8 | 5 | 5 | 9 | 2 |
| Health habits..... | 10 | 5 | 15 | 3 | 1 | |
| Physics and astronomy. | 9 | 16 | 6 | 10 | 13 | 22 |
| Chemistry..... | 2 | 11 | 4 | 9 | 4 | 7 |
| Geology and weather... | 4 | 5 | .. | 2 | 7 | |
| Physiology..... | 5 | 5 | 4 | 6 | 4 | 7 |
| Miscellaneous..... | 7 | 5 | 4 | 4 | 1 | 5 |
| Total..... | 60 | 60 | 50 | 50 | 52 | 52 |

indicates that as we move up to grades 7 and 8 there is a decrease in the number of items in animals, plants, and health habits and an increase in items on physics, astronomy, and chemistry. One illustration from each of the four main divisions of grade 5 of the Coordinated Scales of Attainment is now presented.

¹ Items by permission of Educational Test Bureau, Minneapolis, Minn.

18. A gnawing animal that lives in the water is the 1 mouse 2 shrew
3 muskrat 4 catfish.
45. Plants may grow tall and pale indoors because of lack of 1 water 2 air
3 light 4 soil.
32. There are few school hall accidents if pupils 1 walk fast 2 play tag
3 hurry to classes 4 walk quietly.
58. The amount of electricity a lamp uses is measured in 1 kilowatts
2 money 3 watts 4 volts.

The following illustrations are from grade 8:

38. The green material in plants helps them to 1 breathe 2 hold water
3 make food 4 produce seeds.
41. A body that has fallen from the sky to the earth is a 1 planetoid 2
meteor 3 meteorite 4 nebula.
53. A common chemical change is 1 rain falling 2 evaporation 3 air
circulating 4 burning.

The Stanford Achievement Test,¹ as can be seen from Table 8, covers about the same areas as do the Coordinated Scales of Attainment, but less extensively. The former emphasizes health habits somewhat more in grades 4 to 6 and physics and astronomy somewhat less in grades 7 and 8. This Stanford Achievement Test uses only three choices in its tests, which increases the chances of guessing. Illustrations for grades 4 to 6 appear below:

19. The best cure for fatigue is—1 coffee 2 rest 3 tobacco 1 2 3
:: :: ::
36. The buzz of a fly is made by its -7 feelers 8 wings 9 legs 7 8 9
:: :: ::
30. Never use an electric appliance when—7 standing on a wet floor 7 8 9
8 camping 9 in bed :: :: ::

The following illustrations are for grades 7 and 8:

20. Which has the most valuable fur? 4 the bear 5 the mink 4 5 6
6 the squirrel :: :: ::
42. The boiling point on the centigrade thermometer is 70° 8 100° 7 8 9
9 212° :: :: ::
39. Iron, lime, and phosphorus are examples of 7 minerals 8 pro- 7 8 9
teins 9 enzymes :: :: ::

The Metropolitan Achievement Tests¹ place considerable emphasis on plants and animals and their relations. Note the much greater number of items dealing with astronomy and physics in both the intermediate and advanced batteries. Geology and the weather also have

¹ Items by permission of World Book Company, Yonkers, N.Y.

the usual number of items. Samples from this battery suitable for grades 7 and 8 and lower 9 are:

32. Telephone wires make a humming noise between poles because they—1 have high pitch 2 are stretched taut 3 carry electricity 4 vibrate.
18. When a raccoon goes to sleep in the fall, it—1 propagates 2 estivates 3 hibernates 4 migrates.
41. To burn food the body needs—1 oxygen 2 air 3 carbon dioxide 4 hydrogen.

It is quite evident from the discussion and from the samples of items that the objective of learning and understanding of facts and information acquired in the various sciences is well measured. On the other hand, there are no items or set of items which reflect attainment in the ability to use the scientific method. There are no problems from which to draw correct conclusions and no opportunity to formulate hypotheses. *The test batteries are best when they test scientific information; worst, when they test scientific thinking.*

SPECIAL TESTS FOR SCIENCE

Here are two illustrations of entire tests devoted to the testing of science.

The Cooperative Science Test for Grades 7, 8, and 9 measures many of the objectives set forth on pages 248 and 249. It is divided as shown in the accompanying table.

| Part | Number of items | Time, minutes |
|--|-----------------|---------------|
| I. Facts, Skills, and Application..... | 75 | 40 |
| II. Terms and Concepts..... | 45 | 15 |
| III. Comprehension and Interpretation..... | 30 | 25 |
| Total..... | 150 | 80 |

The items of Part I are taken from problems which arise in everyday living. The facts and skills are usually tested in a meaningful setting. The amount of starch in foods, the superstition involved in touching toads and developing warts, what sleet is, how malaria is carried, what the Milky Way is, why milk is the best food for growing children, how plants are pollinated—these illustrate the rich variety of topics sampled. Two illustrations are:¹

¹ Items used by permission of Educational Testing Service, Princeton, N.J.

12. Children of school age are vaccinated as a protection against
 - 12-1 malaria
 - 12-2 smallpox
 - 12-3 tuberculosis
 - 12-4 scarlet fever
 - 12-5 influenza
52. Of the following substances, which are the hardest?
 - 52-1 iron
 - 52-2 steel
 - 52-3 cement
 - 52-4 diamond
 - 52-5 granite

Part II, Terms and Concepts, recognizes that in terms and concepts frequently are concentrated whole areas of meanings and generalizations. They are of the first importance also in reading scientific material. The following are illustrative terms: "constellation," "architect," "calorie," "disinfectant," "convection," "respiration," "experimentation," "microphone," "oxidation," "mammal," "element," and "battery." Two items are:

11. The plants, animals, and physical world making up the surroundings of man are called his
 - 11-1 adaptation
 - 11-2 heredity
 - 11-3 environment
 - 11-4 circumstances
 - 11-5 vicinity
37. A device which has been used successfully for exploring the ocean at great depths is the
 - 37-1 bathysphere
 - 37-2 hydrosphere
 - 37-3 stratosphere
 - 37-4 vivarium
 - 37-5 depth bomb

11()

37()

Part III, Comprehension and Interpretation, is composed of six paragraphs on science. All of them are rather simply written and contain questions both on the facts of the paragraph and on their application and interpretation. The six paragraphs include one on the best ways to preserve and plant tulip bulbs, one on the importance of the new turnpike from Harrisburg, Pa., to Pittsburgh, and one on the interaction between plants and oxygen. One of the questions asked requires the student to understand what the principal idea of the paragraph is. It is now well recognized that teaching of reading must be done in every

grade and throughout high school. The teaching and understanding of reading materials in science is of the first importance.

Another good test in the field of general science is the Ruch-Popenoe General Science Test.¹ This test, published in 1923, is composed of two parts: Part I, on terms and concepts, and Part II, consisting of drawings with accompanying questions. Part I consists of 50 terms and concepts. It uses the multiple-choice form of testing, with seven choices. These 50 terms sample well the material usually covered in a general science course. Samples of concepts are "oxidation," "pollination," and "ductility." Two illustrations are:

17. The act of transfer of pollen from anther to stigma is called
 pollination reproduction fertilization transpiration mitosis
 adaptation filtration.

46. Glucose is found in large quantities in
 eggs grapes olive oil beefsteak onions rice tapioca.

The second part consists of 20 drawings, with two to five questions asked about each drawing. The questions are either of the completion or short-answer form. Drawings with their appropriate questions delve into such scientific problems as the names of the parts of a flower, the principle of the lever, the mechanical efficiency of pulleys, the lifting power of a pump, and the understanding of the general process of artificial freezing. One illustration will show the general technique.



In this diagram of the digestive tract:

- | | | |
|----------|--------------------------------------|----------|
| <i>a</i> | The small intestine is lettered..... | <i>a</i> |
| <i>b</i> | The œsophagus is lettered | <i>b</i> |
| <i>c</i> | The liver is lettered | <i>c</i> |
| <i>d</i> | The stomach is lettered | <i>d</i> |
| <i>e</i> | The pancreas is lettered | <i>e</i> |

This test has two forms, a reliability of .83, and consumes 40 minutes in the taking. Probably its greatest weakness is its failure to have a reading test of scientific material.

For the high school there is A Test of General Proficiency in the Field of Natural Sciences by Paul J. Burke, one of the Cooperative General Achievement Tests. It also is divided into two parts: Part I, terms and concepts, and Part II, comprehension and interpretation. The time consumed in the actual work of taking the test is 40 minutes. The test is more advanced than those previously described.

¹ Items by permission of World Book Company, Yonkers, N. Y.

Part I asks questions about the meaning of "fossils," "calorie," "abrasive," "ion," "lymph," "wiggler," "the momentum of an object," and so forth. There are 50 items in all. One illustration is:¹

43. A substance which increases the number of hydrogen ions in a solution is known as
- 43-1 a base
 - 43-2 a salt
 - 43-3 a buffer
 - 43-4 an acid
 - 43-5 an alkali

Part II deals with the understanding of paragraphs. Many of the questions ask that the subject apply the principle stated in the paragraph to new illustrations. There are two reading selections and one table dealing with the amount of theoretical horsepower required to raise water to different heights. From this table several problems in physics are constructed. This test covers two areas of general science, biology and physics. Percentile norms are available.

TESTS OF SCIENCES IN HIGH SCHOOL

TESTS OF BIOLOGY

Our best standard tests in biology sample well the information which a student has acquired. Frequently items are so arranged that some reasoning and thinking are required to arrive at the correct answer. In one or two tests, subjects are asked to predict the outcome under the conditions named. In no cases are the reasons required for the conclusion given nor are hypotheses to be formulated for the explanation of facts. Moreover, both the planning of experiments to solve pressing problems and the understanding of the nature of proof are neglected. It is, therefore, well to remember that none of the tests described measure all the outcomes of the teaching of biology. It is always important to ask about any test, "What aspects of biological instruction does this instrument test well?"

The Cooperative Biology Test¹ is one of the better types of standardized tests. Its items, based on information usually taught in biology courses, are so constructed that real thinking is required to answer them correctly. The norms of this test are well established. The reliability as computed by the odd-even technique is .94 and thus is satisfactory. This test is divided into two parts. Part I, which requires 25 minutes of testing time, is composed of 75 items. Many of the items are taken from problems met in daily life. How to get rid of cockroaches, what to worry about in case of termites in the neighborhood, types of insects

¹ Item by permission of Educational Testing Service, Princeton, N.J.

which reduce yield from a hay field, what is the best thing to do about influenza, and what a morning sore throat implies—these are illustrations. Physiological material is emphasized more than morphological. Two samples from Form Q will indicate the manner in which functional information is tested:

28. Which of the following factors is part of the normal environment of deep sea organisms and not of land organisms?
- 28-1 The presence of oxygen
 - 28-2 The presence of mineral salts
 - 28-3 Great pressure
 - 28-4 The presence of natural enemies
 - 28-5 Freezing temperatures 28()
49. A certain species of land plant develops broad leaves which contain chlorophyll. This indicates that this plant
- 49-1 grows best in dry regions
 - 49-2 will grow only on acid soils
 - 49-3 is able to make food from carbon dioxide and water
 - 49-4 is able to survive extreme variations in temperature
 - 49-5 is probably a type of fungus 49()

Part II, which requires 15 minutes of testing time, contains both 24 matching problems, which in most cases involve the understanding of drawings, and also 21 items of the best-answer type. From a drawing of a tooth, for example, the subject has to recognize the various parts of a tooth; from drawings, he must recognize certain types of cells; etc. On the other hand, there are no drawings in the last 21 items. The questions are asked in such a way as to require considerable thought to answer them correctly. One must recognize a disease which can be made less common by a better diet, and from the knowledge that simple animals are better able to regenerate lost or injured parts one must infer that this animal is the starfish. Two illustrations from Form Q are:

36. A single-celled organism is found to have cytoplasm, a nucleus, chloroplasts, vacuoles, and a cell membrane. Which of these indicates that the organism is a plant rather than an animal?
- 36-1 The cytoplasm
 - 36-2 The chloroplasts
 - 36-3 The nucleus
 - 36-4 The cell membrane
 - 36-5 The vacuoles 36()
42. L represents long-haired, which is dominant; s represents short-haired, which is recessive; LL is crossed with ss. The offspring in the first generation will be in the ratio of
- 42-1 $2LL + 2ss$

42-2 4LL

42-3 4ss

42-4 LL + 2Ls + ss

42-5 4Ls

42()

Percentile norms are available for this test.

A second test, the Ruch-Cossman Biology Test, despite its age (1924) is worthy of consideration. The items for this test were selected from examination questions supplied by 126 teachers, who were asked to send in to the investigator copies of the examination questions used during that year. From the 2,000 questions received, the 300 occurring most frequently were selected. These questions were then rated by "68 teachers and 9 authorities." Each question was rated "1" if entirely satisfactory, "2" if partially satisfactory, and "3" if entirely unsatisfactory. Most of the items selected for the test came from those rated "1." The test's reliability ranging in coefficients from .76 to .87 as computed from populations ranging in age from 12 to 28. By combining Form A and Form B into one test a satisfactory reliability of .90 or above was obtained. The probable error of measurement is three points.

The Ruch-Cossman Biology Test is composed of five tests or parts. Test 1 is composed of 40 terms whose correct definitions or meanings appear among seven possible answers. The student is asked about the action of gravitation on roots, about chlorophyll, what mandibles, enzymes, and collar cells are. Two illustrations are:¹

28. The stage of an embryo which most closely resembles a hollow ball is the
 ovum blastula pupa gamete gastrula chrysalis zoöspore.
37. Fehling's solution is a test for
 fats cellulose glucose albumin starch proteins minerals.

Test 2 is composed of 18 incompleated statements which are completed by checking one of three statements

13. The arthropods always possess
 _____ Three distinct body regions
 _____ Two pairs of wings
 _____ Jointed appendages

Test 3 matches 18 names of structures with their positions in four drawings.

Test 4 has two illustrations of the working of Mendelian inheritance.

Test 5 is made up of five paragraphs, each of which has certain key words omitted. The usual difficulty found in marking completion tests is present in this test.

This test samples well the worth-while facts learned in a biology

¹ Items by permission of World Book Company, Yonkers, N.Y.

course in high school. It does not attempt to test a student's capacity to formulate hypotheses, to set up experiments, to test hypotheses, or to reason logically.

TESTS OF CHEMISTRY

The Cooperative Chemistry Test is divided into two parts. Part I, which requires 25 minutes of testing time, contains 56 items constructed in the best-answer manner. With few exceptions the questions require a functional understanding of the chemical terms and processes. The second part contains 39 questions. These two illustrations are from Part I:¹

11. Some paints darken on standing. This is caused by the formation of
- 11-1 ZnS
 - 11-2 ZnSO_4
 - 11-3 PbS
 - 11-4 PbSO_4
 - 11-5 TiO_2
- 11()
31. The catalyst in the contact process affects which of the following changes?
- 31-1 $\text{S} + \text{O}_2 \rightarrow \text{SO}_2$
 - 31-2 $2\text{SO}_2 + \text{O}_2 \rightarrow 2\text{SO}_3$
 - 31-3 $\text{H}_2\text{SO}_4 + \text{SO}_3 \rightarrow \text{H}_2\text{S}_2\text{O}_7$
 - 31-4 $\text{SO}_3 + \text{H}_2\text{O} \rightarrow \text{H}_2\text{SO}_4$
- 31()

The next two illustrations are from Part II:

10. When CO_2 is bubbled into limewater, a white precipitate forms which dissolves upon the further addition of CO_2 . The substance finally remaining in solution is
- 10-1 CaO
 - 10-2 $\text{Ca}(\text{OH})_2$
 - 10-3 $\text{Ca}(\text{HCO}_3)_2$
 - 10-4 CaCO_3
 - 10-5 $\text{Ca}_2(\text{OH})_2(\text{CO}_3)$
- 10()
25. One of the products in the completely balanced reaction between ZnCl_2 and AgNO_3 is
- 25-1 ZnNO_3
 - 25-2 AgCl_2
 - 25-3 2ZnNO_3
 - 25-4 2AgCl
 - 25-5 $\text{ZnAg}(\text{NO}_3)_2$
- 25()

Some of the questions in this test are directly factual. Thus the subject is asked about the facts that vinegar contains acetic acid, that tungsten is used for filaments of ordinary light bulbs, and that stainless steel has chromium in addition to iron and about the formula for heavy water. There are some practical problems such as the selection of

¹ Items by permission of Educational Testing Service, Princeton, N.J.

an instrument to test a storage battery, or to know the contents of baking powder, the reason why most gold in use is alloyed with copper. A great majority of the items would fall under the head "a knowledge of, and ability to use, fundamental tools of chemistry." The authors of the test, Form P, expect to measure five principal types of objectives:

(a) Knowledge and understanding of chemical laws, principles and theories.

(b) Knowledge of, and ability to use, fundamental tools of chemistry.

(c) Understanding and appreciation of applications of chemistry in industrial processes and in daily life.

(d) Ability to perform correctly simple basic calculations in chemical problems involving the application of chemical principles.

(e) Knowledge and appreciation of great chemists and their contributions.

As a whole the test does well what it sets out to do, but probably "tests knowledge of and ability to use fundamental tools of chemistry best of all." Because its items were based on the content of four widely used textbooks in chemistry they cover well the field of traditional chemistry, but not so well the field of modern chemistry.

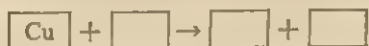
The test has five forms N, O, P, Q, and S and has satisfactory norms. Scaled norms and percentile points are furnished for both public and college preparatory schools. Its reliability is well above .90 when computed from the scores secured from the members of one class. Its correlation with school marks run from .63 to .78. Traxler showed¹ that with intelligence constant the correlation of this test and school marks in independent secondary schools is .64.

Another test is the Columbia Research Bureau Chemistry Test.² It is divided into three parts. Part I consists of 150 items constructed according to true-false principles. Two illustrations are:

25. Carbon monoxide is important in metallurgical industries because it is a reducing agent. ()
44. Sodium nitrate is the only commercially important mineral source of fixed nitrogen. ()

Part II is composed of 22 exercises which deal with the completion and balancing of equations. Examples are:

5. Metallic copper placed in an aqueous solution of silver nitrate



¹ Traxler, Arthur E., "Correlation of Achievement Scores and School Marks," *School Review* (1937) 45: 198-201.

² Items by permission of World Book Company, Yonkers, N.Y.

14. The action of water on phosphorous tribromide



Part III contains 10 problems to be solved. Two items are:

4. Calcium carbonate is acted upon by HCL according to the following equation:
 $\text{CaCO}_3 + 2\text{HCl} \rightarrow \text{CaCl}_2 + \text{CO}_2 + \text{H}_2\text{O}$. Suppose you have 50.0 grams of CaCO_3 to convert into CaCl_2 , what is the minimum amount of HCL you will have to furnish (in grams)? ()
- (Atomic weights: Ca = 40, C = 12, O = 16, Cl = 35.5, H = 1)
8. A gas under a pressure of 800 millimeters of mercury and at a temperature of 27°C . occupies 100 liters. How many liters will the same weight of gas occupy, if the pressure is decreased to 400 millimeters and the temperature raised to 177°C .? ()

The norms are constructed from the scores of some 8,000 high school students in one state. The reliability is reported as .87.

This is an older test (1928) than the Cooperative test described above. It consumes 110 minutes of testing time. In general, the longer the test, the higher is the reliability, but as between these two tests the opposite is true. The employment of the true-false technique is probably a weakness. A worse fault is the great emphasis in this test on factual detail. On the other hand, the balancing of equations and the solution of problems involve discrimination, interpretation of facts, and reasoning.

TESTS OF PHYSICS

Only one test of physics will be described, but others are listed at the end of the chapter.

The Cooperative Physics Test consists of 85 items constructed in such a manner that each item is introduced with an incomplete statement which is followed by five choices, one of which completes the statement. The items of the test are based on the syllabuses of the College Entrance Examinations Committee and the New York Board of Regents. The number of tests on each topic is somewhat proportional to the emphasis given to that topic in the syllabuses. There is general agreement among the reviewers of this test that the majority of the problems involve discrimination, interpretation of facts, and reasoning.

The two parts of the test contain nearly the same number of items and require 20 minutes of working time each. Part I has an irregular arrangement of items dealing with many aspects of physics. The subject must jump from the work that can be done by a 10-horsepower engine, to the pressure exerted on fish under water, to a consideration of the problem of the vector of two forces. The second part is more unified, having 22 items on electricity, 14 on light, and 6 on sound. The following example is from Part I:¹

¹ Items by permission of Educational Testing Service, Princeton, N.J.

18. A stone falls freely from rest. At the end of $\frac{1}{2}$ second its speed is approximately
 - 18-1 8 ft. per sec
 - 18-2 2 ft. per sec
 - 18-3 16 ft. per sec
 - 18-4 4 ft. per sec
 - 18-5 32 ft. per sec
29. The calorie is a unit of
 - 29-1 weight
 - 29-2 temperature
 - 29-3 force
 - 29-4 power
 - 29-5 energy

These examples are from part II:

9. The electric current in a horizontal wire is from south to north. If a compass needle is placed beneath the wire, its N-pole will be
 - 9-1 undeflected
 - 9-2 deflected downward
 - 9-3 deflected upward
 - 9-4 deflected toward the east
 - 9-5 deflected toward the west.
26. The fact that a candle flame gives a continuous spectrum is evidence that it contains
 - 26-1 luminous gases
 - 26-2 unburned gases
 - 26-3 gases of different temperatures
 - 26-4 droplets of warm liquid
 - 26-5 particles of an incandescent solid.

This test is well standardized. It has forms N, O, P, Q, and S available as well as percentile norms both for preparatory schools and public schools. Scaled scores and standard errors of measurement are also provided. The reliability for the 40-minute test is in the neighborhood of .92.

There are three or four minor criticisms of this test. In the first place, 85 items to be done in 40 minutes leaves little time for contemplation. In the second, by arranging the items irregularly in Part I of the test the subject is compelled to shift quickly from one item to another. In the third place, there are some items which merely require identification, a one-step mental process. It is possible that a few more problems which demand reasoned understanding might improve the test.

As a whole, this test satisfies more nearly than any other physics test the criteria of a good test, largely because it tests the understanding of significant facts and principles of physics. This test has a reliability of .92 to .97 and a correlation of .73 with school marks.

INSTRUCTIONAL TESTS IN SCIENCE

Four instructional tests are here described. These are (1) Blaisdell Instructional Tests in Biology, (2) Glenn-Welton Instructional Tests in Chemistry, (3) Glenn-Obourn Instructional Tests in Physics, and (4) Glenn-Greenberg Instructional Tests in General Science. There are 25 units of work in biology and physics and 36 units in chemistry. For each unit of work presumably taught from any standard textbook there is a complete, standardized test composed usually of 25 to 50 items and in some cases a longer over-all test by way of review at the end of a division.

The authors claim, and for the most part justly, that these tests are useful in the following ways:¹

1. to provide information about student achievement on which to base instructional-practices.
2. to diagnose learning difficulties of students and study the nature of their errors separately for each unit of beginning Chemistry.
3. to make frequent inventories of a student's success with a reasonable expenditure of time.
4. to reveal to both teacher and student the outstanding difficulties that students have in learning chemistry as a basis for an intelligent drill program of remedial work through the semester.
5. to investigate problems relating to learning in Chemistry and thus make a beginning in the development of the psychology of chemistry.

If the teacher organized his course in the same manner as the courses from which the instructional tests were constructed, these tests would undoubtedly be of great service. They would almost certainly facilitate the mastery of each unit. Unfortunately for the use of the tests, but fortunately for the development of understanding, much of the good work in science comes in interpreting the environment in which the students find themselves. Units of science instruction arise from students' questions and from the problems they raise. Some points suggested in the tests are omitted from the usual course while others are added. For this reason, instructional tests are not as widely used as the care of their preparation would lead one to expect. The questions and techniques employed in these tests contain materials highly suggestive to that teacher who aims at mastery of the material covered.

¹ Manual for Glenn-Welton Instructional Tests in Chemistry, P III. World Book Company, Yonkers, N.Y. By permission.

SCIENTIFIC THINKING

Scientific thinking is an outcome of every scientific relationship that is perceived, every problem that is exactly solved. It is developed when children are taught to delay their inferences until all the data are in. It is encouraged when a student makes no statement in geometry unless the grounds for his proof are also presented. Wherever critical analyses are made of the facts presented, there scientific method is beginning. Finally, when an individual acquires a mind which is willing to accept the facts and draw his conclusions from them, he has made progress toward scientific thinking.

These characteristics—of perceiving relations in scientific data, of valuing accuracy of result, of withholding inferences until all the data are in or of asking for more data, of demanding grounds for statements, of critically analyzing data present, and of being willing to accept facts and to draw conclusions from them—are well known to all teachers of science. Many of them, however, are too much carried away by the load of detail which their students must master to take the trouble to instruct students in the scientific method. Scientific method has a definite transfer value. Properly developed with one sort of data, broadly illustrated, and contrasted with the method of superstition or common sense, it extends far beyond the biology, physics, or chemistry where it is learned to much broader areas of the social sciences and to thinking in general.

The tests suggested here are pretty largely checks to see if the students are able to apply their scientific method to new situations. The illustration introduced below tests to see if a pupil can draw the right conclusion from rather simple data and then can check the correct reasons.¹

Form 1.3

APPLICATION OF PRINCIPLES

Directions: In each of the following exercises a problem is given. Below each problem are two lists of statements. The first list contains statements which can be used to answer the problem. Place a check mark (✓) in the parentheses after the statement or statements which *answer the problem*. The second list contains statements which can be used to explain the right answers. Place a check mark (✓) in the parentheses after the statement or statements which *give the reasons for the right answers*. Some of the other statements are true but do not explain the right answers; do not check these. In doing these exercises then, you are to place a check mark (✓) in the parentheses after the statements which *answer the problem* and which *give the reasons for the RIGHT answers*.

¹ Smith, Eugene R., Ralph Tyler, *et al.*, *Appraising and Recording Student Progress*, pp. 88-90. New York: Harper & Brothers, 1942. By permission.

In warm weather people who do not have refrigerators sometimes wrap a bottle of milk in a wet towel and place it where there is a good circulation of air. *Would a bottle of milk so treated stay sweet as long as a similar bottle of milk without a wet towel?*

A bottle wrapped with the wet towel would stay sweet

- a. longer than without the wet towel. () a.
 b. not as long as without the wet towel. () b.
 c. the same length of time—the wet towel would make no difference () c.

Check the statements below which give the reason or reasons for your explanation above. Statements in the left column are used in scoring. They do not appear on the test.

| | | |
|----------------------|--|--------|
| Superstition | d. Thunderstorms hasten the souring of milk | () d. |
| Right Principle | e. The souring of milk is the result of the growth and life processes of bacteria. | () e. |
| Wrong | f. Wrapping the bottle prevents bacteria from getting into the milk. | () f. |
| Wrong | g. A wet towel could not interfere with the growth of bacteria in the milk. | () g. |
| Wrong | h. Wrapping keeps out the air and hinders bacterial growth. | () h. |
| Right Principle | i. Evaporation is accompanied by an absorption of heat. | () i. |
| Authority | j. Milkmen often advise housewives to wrap bottles in wet towels. | () j. |
| Unacceptable Analogy | k. Just as many foods are wrapped in cellophane to keep in moisture, so is milk kept sweet by wrapping a wet towel around the bottle to keep the moisture in. | () k. |
| Right Principle | l. Bacteria do not grow so rapidly when temperatures are kept low. | () l. |

A second illustration involving pretty largely the facts learned in science is now given. In this case facts and assumption must be distinguished.¹

Exercise 21

Are you learning to recognize and evaluate assumptions?

A small piece of magnesium will ignite and burn with a bright light in an atmosphere of chlorine gas, leaving white ashes. Bill secured some chemicals which, when mixed together and heated, gave off a colored gas. He collected some of this gas in a bottle. The chemistry teacher gave him a small piece of magnesium. Bill put it in

¹ "The Measurement of Understanding," *Forty-fifth Yearbook of the National Society for the Study of Education*, Part I, pp. 132-134; Chicago: University of Chicago Press, 1946. By permission.

the bottle of colored gas. The magnesium ignited, burned with a bright light, and left white ashes. Bill told his friends that his results conclusively proved that the colored gas was chlorine.

Part 1. Directions: Read each statement below. Is the statement a **FACT**, or is it an **ASSUMPTION**? Place a check mark (✓) in the appropriate column *before* the statement.

Part 2. Directions: Read over again only those statements which you have marked as *assumptions*. Place a check mark (✓) after those **TWO ASSUMPTIONS** which are absolutely necessary in proving that the gas was chlorine. Do not mark more than two.

Statements

| Fact | Assump- tion | | |
|--------------------------|--------------------------|---|-----------------------------|
| <input type="checkbox"/> | <input type="checkbox"/> | a. Chlorine is not the only gas in which magnesium will burn with a bright light and leave white ashes. | <input type="checkbox"/> a. |
| <input type="checkbox"/> | <input type="checkbox"/> | b. The material the chemistry teacher gave him was magnesium. | <input type="checkbox"/> b. |
| <input type="checkbox"/> | <input type="checkbox"/> | c. Chlorine gas is the only gas in which magnesium will ignite. | <input type="checkbox"/> c. |
| <input type="checkbox"/> | <input type="checkbox"/> | d. Chlorine gas is the only gas in which magnesium will ignite, burn with a bright light, leaving white ashes. | <input type="checkbox"/> d. |
| <input type="checkbox"/> | <input type="checkbox"/> | e. Bill mixed and heated some chemicals which gave off a colored gas. | <input type="checkbox"/> e. |
| <input type="checkbox"/> | <input type="checkbox"/> | f. A small piece of magnesium will ignite and burn with a bright light in an atmosphere of chlorine gas, leaving white ashes. | <input type="checkbox"/> f. |
| <input type="checkbox"/> | <input type="checkbox"/> | g. Chlorine gas is the only gas in which magnesium will burn with a bright light. | <input type="checkbox"/> g. |
| <input type="checkbox"/> | <input type="checkbox"/> | h. Bill collected some of the colored gas in a bottle. | <input type="checkbox"/> h. |
| <input type="checkbox"/> | <input type="checkbox"/> | i. The properties of the colored gas in the bottle were the only cause of the magnesium igniting, burning with a bright light, and leaving white ashes. | <input type="checkbox"/> i. |
| <input type="checkbox"/> | <input type="checkbox"/> | j. Bill put a small piece of magnesium in the bottle. | <input type="checkbox"/> j. |
| <input type="checkbox"/> | <input type="checkbox"/> | k. The properties of the colored gas in the bottle were not the cause of the magnesium igniting, burning with a bright light, and leaving white ashes. | <input type="checkbox"/> k. |
| <input type="checkbox"/> | <input type="checkbox"/> | l. The magnesium ignited, burned with bright light, and left white ashes. | <input type="checkbox"/> l. |

Are you learning how to develop a logical proof?

When arguments for or against some proposition are presented in newspapers, magazines, speeches, or textbooks, we often feel that the discussion could have been made more logical. Authors sometimes put in statements that are really unnecessary to prove their point; at other times they leave out important arguments; on still

other occasions they arrange their statements in such poor order that the conclusion does not seem to be based on or to grow out of the arguments.

Part 3. *Directions:* Suppose you were describing this experiment in order to prove that chlorine gas was collected. What are all of the *absolutely* necessary statements in the complete development of the proof? Use as many of the above statements as are necessary and place the letters of these statements in their proper order¹ on the line below. Do not use any unnecessary statements.

Are you learning to support your own conclusions with sound arguments?

Part 4. *Directions:* In Part 3 of this test you presented a logically developed proof which reached the conclusion that the colored gas Bill made must be chlorine. You may or may not believe that it has been adequately proved that the colored gas must be chlorine.

Check the following statement which best represents your own personal opinion as to the nature of the gas.

- _____ a. I believe that the colored gas Bill made was chlorine.
_____ b. I do not believe that the colored gas Bill made was chlorine.
_____ c. I do not believe that it has been adequately proved that the colored gas Bill made was chlorine.

Write out the reasons you have to support your opinion.

Evidence concerning the student's understanding of good and poor analogy, avoiding a repetition of a conclusion and certain other elements of good reasoning may be obtained from an analysis of his responses to test items constructed like one described under Application of Principles, page 263.

Much of the material in Chap. VI of the *Forty-fifth Yearbook of the National Society for the Study of Education* and in Chap. II of Smith and Tyler is concerned with developing procedures to inculcate in children the habit of thinking scientifically. Illustrations of informal tests, which the teacher can utilize or imitate, to measure the progress of students in using the scientific method are there presented.

ATTITUDES AND INTERESTS IN SCIENCE

Attitude, as is pointed out in Chap. 17, consists of a learned tendency, set, or disposition to act favorably or unfavorably toward an object, process, situation, or person. It is not the *habit* of accuracy but the *set* or disposition to be accurate. It is not the *making* of an exact report about an occurrence but the *disposition to make* an accurate report. In

¹ Although the test requests "proper" order, various orders are equally acceptable and the test has been scored in terms of whether all relevant facts and assumptions are included.

most cases, both in interest and attitude, there is a feeling tone which accompanies either the attitude or the interest. "Interest is the pleasant feeling tone which attaches itself either to the activity or to the goal."¹ It is evident, therefore, that attitudes and interests so defined are very difficult to measure. In fact, only indirectly and through inference are we able to get a better understanding of the presence of attitudes or interests.

There are two procedures which could be used to discover the presence of scientific attitudes. The first one would inquire of the students about their interests by means of a questionnaire or self-rating scale. The second would make observations and anecdotal records of those events in which students showed interest or the lack of it. Up to the present time, the second procedure has been most fruitful. In the class itself there are many opportunities for observing activities which reflect students' attitudes. Consider the number of questions asked, the willingness to participate in class demonstrations, the desire to be accurate in reporting, and the inclination to get to the bottom of problems. In all these activities opportunity is offered to observe the results of dispositions. Suppose we add to these opportunities those offered in the home. Accurate reports of electric stoves mended, of pigs raised, of farm machinery put in service, or of animals bred and raised in a scientific manner furnish further indicators of attitude and interests. Books and magazines read are a third source of valuable information. *Science and Invention*, *Popular Mechanics*, and such periodicals contain much material about science, and if a boy reads regularly such a magazine he is reflecting definitely his interest. If all these activities indicate the presence of interest and a scientific attitude, then there is little reason to doubt that it is present.

The best procedure to quantify such attitudes and interests would be to formulate a check list of activities which reflect the presence of attitudes. Each activity should then be given a weight according to the teacher's best judgment as to its value in reflecting an important attitude. The list and the weights would be modified from year to year until a fairly stable form for that community could be achieved. Probably the student should not be aware of the check list for then the "eager beavers" and "teacher pleasers" would be performing these acts but not possessing the attitude. Such a carefully prepared check list could indicate to the teacher whether the student was achieving one of the most important outcomes of science teaching, namely, the scientific attitude.

¹ Jordan, A. M., *Educational Psychology*, p. 155. New York: Henry Holt and Company, Inc., 1942. By permission.

SUMMARY

Five types of objectives for the teaching of science have been described: (1) the learning and understanding of scientific facts, (2) the development of the scientific method, (3) the development in children of habits of healthful living, (4) the development of interest in scientific problems, and (5) the development of the appreciation of the beauties of nature. The greatest success in measurement has been in the learning and understanding of scientific facts. This fact is true of tests suitable both for the elementary and the high school.

In the elementary school, standardized tests of science information have appeared as members of the achievement batteries. Test makers have attempted to place these factual items in a meaningful setting and to include items on healthful living. In some of the tests the sampling of learned scientific facts was quite adequate. In addition, general science tests were developed. These tests tended to cover more thoroughly the areas tested by the upper levels of the achievement-test batteries. They place more emphasis upon the interpretation of facts and upon the drawing of inferences from scientific situations.

At the high school level, tests are constructed for particular subjects such as biology, chemistry, and physics. These tests check the knowledge of facts, of course, but they also set problems which require processes of comparison and inference—in short, of reasoning. These problems also demand the integration of knowledge to answer them correctly.

Tests of the presence and application of scientific thinking were included to indicate the direction that objective tests of this important trait should take. The suggestion was made that check lists might be constructed by the teachers of each community to furnish more quantitative evidence of appreciation in science.

LISTS OF SCIENCE TESTS

I. GENERAL SCIENCE

1. Analytical Scales of Attainment in Elementary Science, grades 5-6, 7-8, 9. 1933. Three levels. One form. Time: 45 minutes. Authors: M. J. Van Wagenen and August Dvorak. Education Test Bureau, Minneapolis, Minn.

2. Applications of Principles in Science, grades 9-12. 1940. One form. Time: 60 minutes. Authors: Committee of Progressive Education Association, Evaluation in the Eight Year Study, Chicago.

3. Cooperative General Science Test, high school. 1939-1947. Forms P, Q, and X. Time: 40 minutes. Author: O. E. Underhill. Cooperative Test Service, New York.

4. General Science Test, National Achievement Tests, grades 7-9. 1936-1939. Two forms. Nontimed (30-45 minutes). Authors: S. R. Powers, Robert K. Speer, Lester D. Crow, and Samuel Smith. Acorn Publishing Co., Rockville Center, N.Y.

5. Science Information Test, grades

4-9. 1937. Two forms. Two levels. Non-timed (about 60 minutes). Elementary, grades 4-6; intermediate, grades 7-9. Author: Everett T. Calvert. Los Angeles, Calif., California Test Bureau.

6. II. A Test of General Proficiency in the Field of Natural Sciences (Cooperative), high school and college. 1947 (revised series). Several forms. Time: 40 minutes. Authors: Paul L. Burke *et al.* Cooperative Test Service, New York.

7. Cooperative Science Test, grades 7, 8, 9. 1941-1947. New forms each year. Time: 80 minutes. Authors: John G. Zimmerman, Richard E. Watson, *et al.*, Cooperative Test Service, New York.

8. Ruch-Popenoe General Science Test, junior high school. 1923. Forms A and B. Time: 40 minutes. Authors: Giles M. Ruch and Herbert F. Popenoe. World Book Company, Yonkers, N.Y.

9. Survey Test of the Natural Sciences, High school and college placement. 1939. Several forms. Time: 40 minutes. Author: Carl P. Swinnerton *et al.* Cooperative Test Service, New York.

10. Examination in General Science, high school level. 1945. Form B. Time: 150 (155) minutes. Authors: Examination Staff of the U.S. Armed Forces Institute. American Council on Education, Cooperative Test Service, New York.

11. McDougal General Science Test, high school. 1941. Forms A and B. Time: 40 (45) minutes. Authors: H. E. Schrammel and Clyde R. McDougal. Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kans.

II. BIOLOGY

1. Cooperative Biology Test, high school. 1939-1947. Forms P, Q, S, and X. Time: 40 minutes. Authors: F. L. Fitzpatrick, S. R. Powers, *et al.* Cooperative Test Service, New York.

2. Ruch-Cossman Biology Test, grades 9-13. 1924. Two forms. Time: 38 minutes. Authors: Giles M. Ruch and Leo

H. Cossman. World Book Company, Yonkers, N.Y.

3. Williams Biology Test, high school. 1934. Two forms. Time: 40 minutes. Authors: John R. Williams and H. E. Schrammel. Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kans.

4. Application of Principles in Biological Science, grades 10-12. 1940. One form. Time: 60 minutes. Authors: Evaluation Staff. Evaluation in the Eight Year Study, Progressive Educational Association. Chicago.

5. Blaisdell Instructional Tests in Biology, high school. 1929. One form. 25 tests in animal, human, and plant biology. One reliable test for each of 25 units of work. Author: J. G. Blaisdell. World Book Company, Yonkers, N.Y.

6. Biology: Every Pupil Test, high school. 1946-1947. New form each year. Author: David B. Davis. Ohio State Department of Education, Columbus, Ohio.

7. Examination in Biology, high school level, grades 10-11. 1945. Form B. Authors: Examination Staff of the U.S. Armed Forces Institute. American Council on Education, Cooperative Test Service, New York.

III. CHEMISTRY

1. Chemistry: Every Pupil Test, high school. 1946-1947. New form each year. Time: 40 (45) minutes. Ohio Scholarship Tests. Ohio State Department of Education, Columbus, Ohio.

2. Columbia Research Bureau Chemistry Test, grades 11-13. 1928-1929. Two forms. Time: 110 minutes. Authors: Eric R. Jette, Samuel R. Powers, Ben D. Wood. World Book Company, Yonkers, N.Y.

3. Cooperative Chemistry Test, high school. 1939-1947. Revised forms P, Q, S, and X. Time: 40 minutes. Authors: S. R. Powers, and Victor H. Noll *et al.* Cooperative Test Service, New York.

4. Cooperative Chemistry Test, Educational Records Bureau Edition, col-

lege preparatory schools. 1941-1943. Three forms. Time: 80 minutes. Norms for preparatory schools only. Authors: Charles L. Bickel, W. Gordon Brown, Robert N. Hilkert, C. S. Hitchcock, and H. H. Loomis. Cooperative Test Service, New York.

5. Examination in Chemistry, high school level. 1944. Form B. Time: 120 (125) minutes. Authors: Examination Staff of U.S. Armed Forces Institute. American Council on Education, Cooperative Test Service, New York.

6. Glenn-Welton Chemistry Achievement Test, high school. 1930-1938. Two forms. Two levels. Test 1, first semester; Test 2, second semester. Time: 71 minutes. Authors: Earl R. Glenn and Louis E. Welton. World Book Company, Yonkers, N.Y.

7. Kirkpatrick Chemistry Test, high school, first and second semesters. 1940-1941. Forms A and B. Authors: Ernest L. Kirkpatrick and H. E. Schrammel. Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kans.

IV. PHYSICS

1. Columbia Research Bureau Physics Test, grades 11-14. 1926. Two forms. Time: 90 minutes. Authors: Herman W. Farwell and Ben D. Wood. World Book Company, Yonkers, N.Y.

2. Cooperative Physics Test, revised series, high school. 1939-1947. Forms P, Q, S, and X. Time: 40 minutes. Machine scorable. Used at end of 1 or of 2 semesters. Author: H. W. Farwell. Cooperative Test Service, New York.

3. Cooperative Physics Test, Educational Records Bureau Edition, college preparatory schools. 1941-1943. Forms ERB-R, ERB-S, and ERB-T. Time: 80 minutes. Authors: Russell S. Bartlett, Lester D. Beers, Winston M. Gottschalk, Robert G. Poland, and Alan T. Waterman. Cooperative Test Service, New York.

4. Fulmer-Schrammel Physics Test, high school. 1934. Two forms. Two parts. Test I, mechanics; Test II, heat, magnetism, electricity, and sound. Time: 40 minutes. Authors: V. G. Fulmer and H. E. Schrammel. Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kans.

5. Glenn-Obourn Instructional Tests in Physics, high school and college. 1930. Twenty-five complete tests, one for each topic. Authors: Earl R. Glenn and Ellsworth S. Obourn. World Book Company, Yonkers, N.Y.

6. Physics: Every Pupil Test, high school. 1946-1947. New form each year. Time: 40 (45) minutes. Author: Darwin J. Kimble. Ohio State Department of Education, Columbus.

QUESTIONS AND EXERCISES

1. List five of the objectives used by teachers of science. Which of these have well-constructed tests for their measurement? Why has the measurement of method been so retarded?

2. What are the leading characteristics of the scientific method? Describe in some detail a test which attempts to measure aspects of the scientific method.

3. What aspects of the scientific method are measured in such an instrument as the Cooperative Chemistry Test?

4. Compare the science tests of the Coordinated Scales of Attainment with

those of the Stanford Achievement test as to (a) method of construction, (b) and coverage of scientific facts. In what respects are they alike?

5. Compare the contents of the science tests occurring in each test battery. Which seem to you the best?

6. Describe in some detail the Cooperative General Science Test. How do you justify a test of science reading in such a test? Why is the measurement of terms and concepts important? What are the limitations of the content in such a test of general science?

7. Do you think that the Cooperative

Biology Test applies facts and principles to the solution of practical problems? Illustrate. Should the records from such a test be used to influence school marks? Why?

8. Name two other tests of biology and describe one of them.

9. How does the test of chemistry cast its problems in a functional setting? What are the principal types of objectives of the Cooperative Chemistry Test? Compare with the Columbia Research Bureau Chemistry Test in the types of material covered and the manner of item construction.

10. Why is there a new interest in physics at the present time? What type of problems are included in the Cooperative Physics Test? Does it test a student's ability to formulate hypotheses or to give reasons for an inference?

11. Show how instructional tests at the end of each unit of work might be used in science. What desirable uses are described? What limitations are there to the use of such tests?

12. Set up a plan for testing the development of attitudes and interests in science.

BIBLIOGRAPHY

CURTIS, DWIGHT K.: *The Contribution of the Excursion to Understanding*, doctor's dissertation, State University of Iowa, 1942.

DAVIS, IRA C.: "The Measurement of Scientific Attitudes," *Science Education* (1935) 19:117-122.

DIAMOND, LEON N.: "Testing the Test Maker," *School Science and Mathematics* (1932) 32:490-502.

EDUCATIONAL RECORDS BUREAU: "Some Data on the Difficulty and Validity of the Cooperative Tests in Biology, Chemistry, and Physics, Forms ERB-R," in 1941 *Achievement Testing Program in Independent Schools and Supplementary Studies*, Educational Records Bulletin No. 33. New York: Educational Records Bureau, 1941.

FLANAGAN, JOHN C.: *The Cooperative Achievement Tests: A Bulletin Reporting the Basic Principles and Procedures Used in the Development of Their System of Scaled Scores*. New York: Cooperative Test Service of the American Council on Education, 1939.

Forty-fifth Yearbook of the National Society for the Study of Education, Part I, "Measurement of Understanding." Chicago: University of Chicago Press, 1946.

FRUTCHHEY, FRED P.: "Illustrative Test Exercises in High School Chemis-

try," *Educational Research Bulletin* (1937) 16:122-26.

GRAY, H. A.: "Approach to the Measurement of Biological Attitudes and Appreciations," *Journal of Educational Research* (1934) 28:25-29.

GREENE, HARRY A., ALBERT N. JORGENSEN, and J. RAYMOND GERBERICH: *Measurement and Evaluation in the Secondary School*, Chap. XIX. New York, Longmans, Green & Co., Inc., 1943.

HAWKES, H. E., E. F. LINDQUIST, and C. R. MANN, (eds.): *The Construction and Use of Achievement Tests*. New York: Houghton Mifflin Company, 1936.

HOFF, A. G.: "A Test for Scientific Attitude," *School Science and Mathematics* (1936) 36:763-770.

NOLL, VICTOR H.: *The Teaching of Science in Elementary and Secondary Schools*, Chap. III. New York: Longmans, Green & Co., Inc., 1939.

ODELL, C. W.: *Educational Measurements in High School*, Chap. VIII. New York: Appleton-Century-Crofts, Inc., 1930.

Redirecting Science Teaching in the Light of Personal-Social Needs, A Report under the Sponsorship of the American Council of Science Teachers in Cooperation with Nine National Societies of Science Teachers of the N.E.A., 1942.

"Science," Vol. IV, *Proceedings of the*

Workshop in General Education. Chicago: University of Chicago Press, 1940.

Science in General Education, Report of the Committee on the Function of Science in General Education, Commission on Secondary School Curriculum, Progressive Education Association. New York: Appleton-Century-Crofts, 1938.

SMITH, Eugene R., RALPH TYLER, et al.: *Appraising and Recording Student Progress*. New York: Harper & Brothers, 1942.

ZAPF, ROSALIND M.: "Superstitious Beliefs," *School Science and Mathematics* (1939) 39:54-62.

CHAPTER 11

Measurement of Business Education

OBJECTIVES IN BUSINESS EDUCATION

When courses in business were first established they were directly related to job preparation. The school was attempting to prepare students who were leaving school early for immediate entry into remunerative occupations. Stenographers, typists, and bookkeepers were needed by the business world. This need was at that time being met by private colleges. The demands and needs of the time led to the introduction of practical courses in business in the high school.

During the last twenty-five years a new impetus has been introduced into business education. Since the publication of *Four Money's Worth*¹ in 1927, it has become clear that the consumer also needs some training in business.² Moreover, school administrators were wondering if there were not some cultural values in these business courses which would be of use to the general student. Gradually, then, there have grown up these two major objectives in business education:

1. To prepare students for immediate jobs through such courses as stenography, bookkeeping, and typing, and in this connection also to help them to (a) become aware of the way business is conducted so that their school subjects will be immediately functional, and (b) become aware of opportunities in clerical work at a higher level, such as secretarial work, as well as of those activities which require technical training.

2. To make of every individual an intelligent consumer of the services of business by acquainting him with the fundamental principles on which business is based. Here the major emphasis will be upon consumer education.

In Division 2, emphasis will be placed on business law, economic geography, and general business. Topics such as advertising, banking, budgets, insurance, taxes, and a host of others which bear directly on the consumer are the ones to be studied.

¹ Chase, Stuart, and F. J. Schlink, *Four Money's Worth*. New York: The Macmillan Company, 1927.

² See Tonne, Herbert A., *Consumer Education in the Schools*, especially Chap. 8. New York: Prentice-Hall, Inc., 1941.

PROBLEMS OF TESTING

From the outline of the purposes and objectives for business education just made, it is immediately apparent that the testing of the outcomes may also be divided into two parts. In one case, habit formation and skills are to be measured; in the other, understandings, comprehension, and information are the major considerations.

CLERICAL TESTS

If we place stenography, bookkeeping, typing, filing, comptometer work, and secretarial duties under the heading "Clerical," then our major problem is to set forth tests of (1) clerical aptitudes, and (2) clerical achievement. In the recent emphasis upon guidance the measurement of clerical aptitude has achieved an important place.¹

TESTS OF CLERICAL APTITUDES

Among the earlier tests of clerical aptitude was the Minnesota Vocational Test for Clerical Workers whose title has now been shortened to the Minnesota Clerical Test. This test consists of (1) 200 sets of numbers, 100 sets of which are the same and 100 different, and (2) 200 sets of names, 100 sets of which are the same and 100 different. The numbers range from 3 digits to 12 digits and the names from 7 to 16 letters.

Here are some sample sets of numbers:²

- | | |
|----------------------------|--------------------------------|
| 121. 46273—46273 | 126. 627152637490—627152637490 |
| 122. 629—620 | 127. 73526189—73526189 |
| 123. 7382517283—7382517283 | 128. 5372—5392 |
| 124. 637281—639281 | 129. 63728142—63728124 |
| 125. 2738261—2728261 | 130. 4783946—4783046 |

Ten items of checking names are:³

121. Bob Fairbanks—Bob Fairbanks
122. Denton Products—Denten Products
123. Wells Dickey Co.—Wells Dickey Inc.
124. S. N. Jonas—S. N. Jonus
125. Warren Co.—Warren Co.
126. Kelly Transfer—Kelly Transfer
127. S. Karpen & Brothers—S. Karpen & Brothers
128. A. J. Drexel—A. J. Drexel
129. C. H. Salmon—S. H. Salmon
130. H. Simons Lbr. Co.—H. Simons Lbr. Co.

¹ See Bingham, Walter Van Dyke, *Aptitudes and Aptitude Testing*, Chaps. XII, XIII, pp. 322-329. New York: Harper & Brothers, 1937.

² Andrew, Dorothy M., Donald G. Paterson, and Howard P. Longstaff, *Minnesota Clerical Test*, New York: The Psychological Corporation, 1933 and 1946. Items by permission.

From the inspection of these samples it is clear that this is a test of perceptual discrimination. The short form of 200 items for each test takes 15 minutes of working time and the long form, which is twice as long, about 30 minutes.

The reliability of this test is about .90. Its validity has been studied in detail. The test has correlated from .54 to .64 with supervisors' ratings of achievement; and it correlates well with other measures of clerical achievement. Name checking correlates with the speed of reading about .45 and with spelling .65; while number checking correlates with arithmetic computation about .51. Its correlation with intelligence is low .23. Critical reviewers of the test state that it is a usable test for selecting promising clerical workers and is a satisfactory instrument for picking out students for clerical training. Its use for over 16 years for these purposes further attests its validity. Criticism is aimed only at its simplicity for it does not test the more complex functions involved in the upper levels of clerical work.

Separate percentile norms are available for men and women in a variety of clerical occupations such as stenography, office machines, clerks, bookkeepers, and accountants, routine clerical workers, etc.

Stenographic Aptitude Tests

A good example of a more specialized aptitude test is the E.R.C. (Educational Research Corporation) Stenographic Aptitude Test. This test, whose author is Walter L. Deemer, consists of five parts:

1. Speed of writing. The subject copies the Gettysburg Address. He writes as fast as he can, but his writing must be legible.
2. Word discrimination. The subject must distinguish between the right use of "current" and "currant," "advice" and "advise," "illusion" and "allusion," "base" and "bass" when used in sentences. There are 34 pairs of words. Moreover, 16 samples of choices between three words in sentences are present in the tests. Illustrations are "writes," "rights," and "rites"; "sight," "site," and "cite"; etc.
3. Phonetic spelling. Fifty phonograms must be spelled out correctly. Here are a few samples: injer, kawf, awt, skeem, hoom.
4. Vocabulary. There are 50 words in short sentences to be defined by choosing from five others the meaning of the word in question. For example, a *fitch* of bacon is to be defined.
5. Dictation. Sentences are dictated at a specified rate.

The reliability is not reported. The author of the test believes that since the validity has been proved satisfactory the reliability must be. However, this is a fallacy because the reliability would show whether further improvement were necessary. If the reliability were .75, considerable improvement would be possible. If it were .93, hardly any more improvement could take place.

Its validity has been well established by correlating it with shorthand achievement ($r .65$) and with accuracy of transcription of material ($r .70$). The test is more exactly a shorthand test than one of stenography since it omits several aspects of stenography. Giving and scoring the test offers a few difficulties. The material for dictation must be given at a defined rate which takes practice to administer correctly. The scoring is tedious because the scorer must count the syllables omitted, inserted, or substituted. Furthermore, the test's efficiency has been demonstrated in grades 11 and 12 but not in secretarial schools. One of its most unique characteristics is a table of predictions. Subjects with scores from 345 to 245 (the subjects within a moderate range) have the scores they will most probably achieve after the passage of two years.

There are several other tests of stenographic aptitude. Three of these will be mentioned briefly. The Turse Shorthand Aptitude Test has seven divisions:

1. Stroking—speed of drawing short lines
2. Spelling—select one or none of three words (45 words)
3. Phonetic association—serten, setl, eksit (60 associations)
4. Symbol transcription—substitution of symbols for letters (six sentences)
5. Word discrimination—select correct word from four, to make good sentence—"Our public schools are founded on democratic (1. principles, 2. principalships, 3. principals, 4. principalities)"
6. Dictation, timed—speed of legible handwriting
7. Word sense (60 words)—phonetic words placed at strategic points in a paragraph

This test is well constructed and standardized and has had considerable use. The two other tests deserving of mention are the Stenographic Aptitude Test (Bennett) and the Detroit Clerical Aptitudes Examination. Critical evaluations of many of these tests and of those which follow in this chapter appear in the *Nineteen Forty Mental Measurement Yearbook* and the *Third Mental Measurement Yearbook*.

CLERICAL ACHIEVEMENT TESTS

Achievement in Stenography

The construction of achievement tests in stenography has been stimulated both by the schools where good standards of teaching were in effect and by businessmen who wished to employ competent stenographers. More lately workers in the Army developed what were usually designated as "examinations" which usually required more time for their administration. An example of the first type is the Turse-Durost

Shorthand Achievement Test; of the second, Stenographic Test, United-NOMA Business Entrance Test; of the third, Examination in Gregg Shorthand.

In all these tests, measures of actual performance play a prominent part. This result is achieved in a variety of ways. Printed words are reproduced in shorthand, shorthand is transcribed into longhand, and sentences in shorthand are to be completed by a choice from several printed words. In some tests a printed article of two or three hundred words is to be written in shorthand outlines on lines above the print, left for that purpose. In one or two tests, syllabication, English, and word usage are added. But in all these tests actual dictation is taken and transcribed.

One example is the Hiatt Stenography Test (Gregg) which includes many of the procedures just described. This test is divided into five parts:

1. Fifty printed words to be reproduced in shorthand
2. Forty shorthand symbols to be transcribed
3. Twenty sentences written in shorthand the completion of which is contained in four printed words
4. An article of 200 printed words, the shorthand outlines to be written above each word on a line left for that purpose
5. Letter dictation (3 minutes) and longhand transcription

Norms are available based on testing 5,296 students in 358 schools after a 1-year course. The reliability is low, .75. Some of the shorthand outlines used for correcting are hard to read, and the directions could be a little clearer.

Another achievement test suitable for the first year of stenographic work is the Examination in Gregg Shorthand. Measures of achievement are secured in three sections.

Section A. 175 printed words and phrases to be written in shorthand.

Section B. Shorthand reading test. The subject transcribes into longhand 300 words.

Section C. Three letters are taken at three different rates of speed: (1) 50 words per minute, (2) 60 words per minute, and (3) 70 words per minute. The rate is controlled by printed material which is marked for timing.

This test, printed in 1944, has no study of reliability or validity that the author has seen up to the present time (1951). However, percentile norms are furnished to the purchasers and practically all the major principles contained in the Gregg manual are contained in the test. Its further use is recommended.

When we turn to the testing of sufficient proficiency for entrance into business, the Stenographic Test, United-NOMA Business Entrance

Tests come immediately to mind. In these tests office managers and competent teachers have combined their efforts to simulate actual office conditions. They have made the test long enough to ensure ample coverage of the skills involved in a realistic office situation.

In this test 30 minutes are given over to dictation, with 5 minutes allowed for extra dictation. There are also allowed 90 minutes for transcription. Nine letters are to be transcribed in mailable form along with straight matter to be typed in the form of a first draft. There is a new edition each year. Percentile norms for the year are furnished schools and business firms. Its reliability is adequate, .90. Some forms have been tried on high school graduates and on those who are regularly employed as typists. The high school students were more apt to become confused during the latter part of the test. Some of them failed to finish the long assignment or else jumbled their work. It will be remembered that these tests are given at regular times only under standard conditions by designated testers.

Achievement in Typing

Like achievement in stenography there are two types of tests: one of these indicates progress toward a less ambitious goal after studying the subject for a year or two; the second indicates an achievement sufficiently advanced for the subject to enter directly into a business office. Representing the first type might be mentioned the Commercial Education Survey Tests. These tests illustrate well the general trend of achievement tests in typing. They are divided into (1) junior typewriting, first year, 95 minutes; and (2) senior typewriting, second year, 120 minutes. The test for junior typewriting is composed of five tests:

Test I. Standard stroking test

Part A. 411 words, 73 per cent from Horne's list of 1000 most common words, 5 minutes

Part B. 407 words, 70 per cent from Horne's list, 5 minutes

Test II. Business-letter test—following instructions in writing a standard business letter, 25 minutes

Test III. Completion test—25 uses of parts of the typewriter, 15 minutes

Test IV. A placement test—mechanics involved in placing a poem on a page

Test V. Centering test—names of twelve of Shakespeare's plays to be typed on a page

The senior test uses the first three tests and adds the typing of a table and a rough-draft test. Its letter to be copied is longer than the one used in the junior test. The scoring is quite typical of the way in

which typewriting tests are scored. If, in the Standard stroking test, 200 words are typed per minute without an error the score is 200. If a word of five letters is omitted then five strokes are subtracted. This would mean one a minute, so the score would be 199. For each error 10 is subtracted from the total strokes per minute, etc. Thus the score is dependent on rate and accuracy.

The test for entrance into business is the Typing Test, United-NOMA Business Entrance Tests. The description of its parts will show that it is not radically different from the Commercial Education Survey Test:

1. Typing a corrected rough draft
2. Setting up a letter from a running copy
3. Simple tabulation on a form
4. Simple tabulation on a plain sheet of paper
5. Typing a form letter with parts to be filled in

Like the preceding test it is scored for (1) form and arrangement of typed matter, (2) accuracy, (3) time consumed, and (4) ability to follow instructions. The reliability is estimated to be .90. Composite total scores include both speed and accuracy. Separate norms for these two factors might be useful under certain conditions. Its norms are percentile scores computed for the year of the testing. These are sent to the teachers and to employers. The tests are administered under standard conditions and sent to a central office for correction. Certificates of proficiency are sent to those who satisfy certain minimum requirements.

Most of the other tests which are now listed are constructed much like the two just described.

LISTS OF TESTS IN STENOGRAPHY AND TYPEWRITING

I. STENOGRAPHIC APTITUDE TESTS

1. Stenographic Aptitude Test, grades 9-16. 1939. One form. Author: George K. Bennett. No validity coefficient for entire test. Psychological Corporation, New York.

2. Turse Shorthand Aptitude Test, grades 8-10. 1940. One form. Time: 45 minutes. Author: Paul L. Turse (see text). World Book Company, Yonkers, N.Y.

3. E.R.C. (Educational Research Corporation) Stenographic Aptitude Test, grades 9 and over. 1944. Time: 33 minutes. Author: Walter L. Deemer (see text). Science Research Associates, Chicago.

4. Minnesota Clerical Test, grades

8-12 and adults. 1933-1946. One form. Time: 35 minutes. Authors: Dorothy M. Andrew, Donald G. Paterson, and Howard P. Longstaff (see text). Psychological Corporation, New York.

5. Detroit Clerical Aptitude Examination, high school. 1937-1944. One form. Time: 30 minutes. Authors: Harry J. Baker and Paul L. Voelker. Public School Publishing Company, Bloomington, Ill.

II. STENOGRAPHIC ACHIEVEMENT TESTS

1. Examination in Gregg Shorthand, first year high school. 1944. Form B. Time: 120 minutes. Authors: Examination staff of the U.S. Armed Forces Institute. Cooperative Test Service, New York.

2. Hiatt Stenography Test (Gregg), high school. 1938-1939. Forms B and C. Two levels. Time: 40 minutes. Authors: Victor C. Hiatt and H. E. Schrammel (see text). Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kans.

3. SRA Dictation Skills, high school and adults. 1947. Six 12-inch records: two for accuracy, four for speed. Authors: Marion W. Richardson and Ruth A. Pedersen. Science Research Associates, Chicago.

4. Stenographic Test, United-NOMA Business Entrance Tests, schools and industry. New form each year. Authors: Joint Committee on Tests, United Business. Educational Association and NOMA. National Office Management Association, New York.

5. Turse-Durost Shorthand Achievement Test, Gregg dictation, 1-2 years in high school. 1941-1942. Time: 60 minutes. Authors: Paul L. Turse and Walter N. Durost. World Book Company, Yonkers, N.Y.

6. Blackstone Stenographic Proficiency Tests, commercial schools or business firms. One form. Time: 50 minutes. Author: E. G. Blackstone. Psychological Corporation, New York.

III. ACHIEVEMENT TESTS OF TYPEWRITING

1. Examination in Typewriting, first and second years high school. 1944. One form. Two levels. First year secondary school, 130 minutes; Second year secondary school, 115 minutes. Examination Staff of U.S. Armed Forces Institute. Cooperative Test Service, New York.

2. Kauzer Typewriting Test, high school. 1934. Three levels: first semester, second semester, and fourth semester. Time: 15-25 minutes. Authors: Adelaide Kauzer and H. E. Schrammel. Bureau of Educational Measurement, Kansas State Teachers College, Emporia, Kans.

3. Typing Test—United-NOMA Business Entrance Tests, school and industry. 1939-1947. New form each year. Authors: Joint Committee on Tests, United Business Educational Association and NOMA. National Office Management Association, New York.

4. Commercial Education Survey Tests, high school. One form. Two levels. Junior typewriting, first year, 95-105 minutes; senior typewriting, second year, 120-130 minutes. Author: Jane E. Clem. Public School Publishing Company, Bloomington, Ill.

BOOKKEEPING TESTS

The objectives in the teaching of bookkeeping are of a practical nature. They are aimed directly at vocational competence. Tests and measures give us an awareness of progress, or the lack of it, toward an ability to make accurate records of the financial transactions of a firm or business. The tests may be divided into those mainly aimed at progress in the school and those which indicate a readiness for entrance into business.

Among the former of these the Examination in Bookkeeping and Accounting is one of the newest and most complete.¹ It now has a test for the first year of bookkeeping and one for the second year. The test for the first year is divided into four parts whose purposes are as follows:

1. Section A tests knowledge of important accounting terms, facts, and principles. This division contains 45 items arranged so that one of four choices is correct. The subject is asked to understand such terms

¹ Items by permission of Educational Testing Service, Princeton, N.J.

as "general ledger," "budget," "drawee," "single proprietorship," "debtor," "creditor," "petty cash fund," "net profit," "gross profits on sales," "net worth," "net loss," "operating expense," etc.

2. Section B tests understanding of the method of adjusting and closing certain accounts. The directions say: "For each of the entries listed below, decide which account should be debited and which should be credited. Show your choice in each case by writing the letters of the accounts in the proper spaces on the answer sheet." The answer sheet is separate from the test.

Accounts

| | |
|--------------------------|---|
| A. Bad debts | I. Profit and loss summary |
| B. Delivery equipment | J. Proprietor's drawing account |
| C. Depreciation expense | K. Purchases |
| D. Expired insurance | L. Reserve for bad debts |
| E. Interest income | M. Reserve for depreciation of delivery |
| F. Interest receivable | N. Sales equipment |
| G. Merchandise inventory | O. Store supplies |
| H. Prepaid insurance | P. Store supplies used |

Example

To record the insurance expired: *Expired insurance* is debited *Prepaid insurance* is credited. Therefore *D* has been placed in the debit column and *H* in the credit column. Look at the answer sheet to see how this has been written.

Answer Sheet

| Db | Cr |
|----------|----------|
| <i>D</i> | <i>H</i> |

Ten statements are to be analyzed in the same way. The two following items are examples:

47. To record the ending merchandise inventory.
51. To record interest accrued on notes receivable.

3. Section C tests skill in analyzing and recording bookkeeping entries in books of original and final entry.

Directions: Assume you are the bookkeeper for William Lane, a lumber merchant. In your answer booklet you will find sections of the following:

| <i>Journals</i> | | <i>Ledgers</i> | |
|-----------------------|------|----------------------------|------|
| | Page | | Page |
| Sales Journal | 2 | General Ledger | 4&5 |
| Purchases Journal | 2 | Accounts Receivable Ledger | 6 |
| General Journal | 2 | Accounts Payable Ledger | 6 |
| Cash Receipts Journal | 3 | | |
| Cash payments Journal | 3 | | |

Step I. Record the following transactions in the proper books of original entry, which are on pages 2 and 3 of your booklet.

Then there follow 13 transactions, dated between August 1 and 31, of which the three following are examples:

August 2. He paid \$120 cash for August rent.

August 14. Sold lumber on account to F. C. Mann, 406 Maple Ave. City. \$600; terms, 2/10, n/30

August 31. Received \$750 from cash sales of lumber, August 1 to 31.

The test then continues as follows:

Step II. Post the journal entries to the proper ledger accounts on pages 4-6 of your answer booklet. The student is warned to (a) post the individual entries to the correct accounts (b) total and rule the proper journals and (c) post the proper journal column balances to the correct accounts.

4. Section D tests skill in preparing a ten-column work sheet.

Directions: On page 7 of your answer booklet you will find a ten column work sheet which you are to complete. The account names and the trial balance amounts are listed on the work sheet. These accounts have *no* connection with the accounts used in Section C. The necessary information for the adjustments is given below on this page.

In preparing the work sheet, you are to:

- A. Enter the necessary adjustments in the "adjustments" column of the sheet.
- B. Complete the other columns of the work sheet in proper form.
- C. Make your entries on the work sheet neatly and in the proper spaces. Be sure to find the net profit or loss and to show all column totals.

Items to be adjusted consist of a changed merchandise inventory, estimated loss from bad debts (\$20), depreciation on office equipment (\$10), accrued salaries payable, etc. The work sheet is to be adjusted after these entries are made.

While there are no reliability and validity studies of this test, it measures well the ordinary procedures used in bookkeeping. Its length (time, 3 hours) may be necessary to measure actual performance.

The second test—Bookkeeping Test, United-NOMA Business Entrance Tests—is, as its name implies, meant to provide information about the proficiency of an individual as a guide to immediate employment. It was constructed, as were the other tests of this series, by a committee representing both the United Business Education Association and the National Office Managers Association. Fitness for immediate employment is indicated by (1) the understanding of the principles and practice of bookkeeping, (2) ability to follow instructions, and (3)

neatness. The test involves (1) a correction of the incorrect entries made in a cash book and journal, (2) correction of the incorrect postings which entails a new trial balance in the general ledger, etc. Some students argue that correcting errors is more like accounting than bookkeeping. The authors, however, argue that "if he can locate and *correct* inaccuracies, that is proof that he can also do the original work."¹

This test has an estimated reliability of .90. The scoring of the test is not entirely objective since it must be rated for neatness on a 10-point scale. From the results of this test certificates are issued. From the standpoint of the teacher this test is of little value except in a general way because the test is administered by experts and scored in a central office.

Several other tests of bookkeeping appear in the following list:

LIST OF TESTS OF BOOKKEEPING

1. Shemwell-Whitcraft Bookkeeping Test, high school, first and second semesters. 1937-1938. Two forms. Two levels. Time: 40-45 minutes. Authors: E. C. Shemwell, J. E. Whitcraft, and H. E. Schrammel. Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kans.

2. Examination in Bookkeeping and Accounting, high school. 1944-1945. One form. Time: 180-190 minutes. Two levels, first year secondary school (1944) and second year secondary school (1945). Section A, knowledge of important accounting terms, facts, and principles, 40 minutes; Section B, understanding of the method of adjusting and closing certain accounts (credit or debit, selecting proper accounts in double entry bookkeeping) 20 minutes; Section C, skill in analyzing and recording bookkeeping entries in books of original and final entry, 75 minutes; Section D, skill in preparing a 10-column work sheet, 45 minutes. Authors: Examining Staff of U.S. Armed Forces Institute. Cooperative Test Service, New York, or Science Research Associates, Chicago.

3. Bookkeeping Tests, State High

Schools Tests for Indiana, first, second, and fourth semesters. 1942-1945. New forms scheduled for each year. Time: 40-45 minutes. Authors: M. E. Studebaker, B. M. Swinford, V. H. Carmichael, F. R. Botsford, and R. Burkheart. State High School Testing Service, Purdue University, Lafayette, Ind.

4. Breidenbaugh Bookkeeping Tests, high school. 1936. One form. Four levels. Single-proprietorship high school bookkeeping course. Test 1, first half of course, nontimed (50-60 minutes); test 2, first half of course, nontimed (50-60 minutes); Test 3, second half of course, nontimed (50-60 minutes), Test 4, second half of course, nontimed (100 minutes). Journalizing, adjustments, balance sheet, statement of profit and loss, closing entries, and worksheet. Author: V. E. Breidenbaugh. Public School Publishing Company, Bloomington, Ill.

5. Bookkeeping Test, United-NOMA Business Entrance Tests, school and industry, 1939-1947. New form each year. One form. Time: 120-130 minutes. Authors: Joint Committee on Tests, United Business Educational Associa-

¹ See *Third Mental Measurements Yearbook* (Oscar K. Buros, ed.), Item 368. New Brunswick, N.J.: Rutgers University Press, 1949.

tion and NOMA. National Office Management Association, New York.

6. Elwell-Fowlkes Bookkeeping Test, high school. One form. Two levels, to be used at end of first and second semester's work. Time: 60 minutes. Measures gen-

eral theory, journalizing, adjusting entries, closing the ledger, and preparing statements. Tests have considerable diagnostic value. Authors: F. H. Elwell and J. G. Fowlkes. World Book Company, Yonkers, N.Y.

There are two other types of work which might be classified as dependent on skill: filing and machine calculation. In each of these areas satisfactory tests have been constructed by the testing committee of United Business Educational Association and National Office Management Association. Their names are (1) Filing Test, United-NOMA Business Entrance Tests, 1939 1947, and (2) Machine Calculation, United-NOMA Business Entrance Tests, 1939 1947. For a complete score in each of these tests their test scores are combined with those of the Business Fundamentals and General Information Test which is described in the next section.

CONTENT TESTS

Under content tests are included:

1. General tests of business information
2. Business English
3. Commercial or business arithmetic
4. Commercial law
5. Economic geography
6. Interest in business

Several aspects of bookkeeping and accounting would also fall under this heading.

Under Item 1 are usually included tests of information which workers in a business office need. Spelling, punctuation, elementary arithmetic, and some knowledge of current events are included. The United-NOMA series of Business Entrance Tests includes such a test in the requirements for certificates in typewriting, stenography, bookkeeping, etc. An illustration of a somewhat different test is the General Test of Business Information (see list) which is suitable for grades 9 to 16. This test includes questions about consumer business education. It asks about the construction of notes and drafts, about buying practices, and about endorsements of notes and drafts. The subject answers questions about the meaning of such terms as "C.O.D." and about the frequency of inventorying personal property. The test claims to cover "the minimum essentials of consumer business information that a high school or college student should possess." There is also some opportunity for diagnosing the results.

The reliability of this test is indicated by a coefficient of .91. Its validity was checked against the subject matter contained in textbooks and syllabuses and by submitting such items to critics in the field.

A second test, very different in nature, is the Business Fundamentals and General Information Test of the United-NOMA Business Entrance Series. This test is not intended for diagnosis and remedial treatment but to indicate proficiency in business. It tests grammar, punctuation, and spelling along with fundamentals in arithmetic and general information usually accumulated from listening to the radio and reading the newspapers. Its reliability is estimated from a previous test made after the same manner and having reliabilities indicated by coefficients ranging from .75 to .84. Its validity is assured by the intimate acquaintance with the field of its constructors, who are a combination of teachers and employers of office workers. No careful study has been made of the correlation between success on this test combined with a test of skill (stenography, typewriting, etc.) and subsequent success in an office.

As for business English, one of the needed tests constructed by the Examination Staff of the Armed Forces Institute is Examination in Business English at the high school level. It is a test of considerable length (testing time, 2 hours) which offers an opportunity to cover the topic thoroughly. There are five sections:

Section I. The selection of misspelled words from a list of 100 words essential to ordinary business communication.

Section II. Word usage—25 pairs of words frequently confused in business, *e.g.*, "principal" and "principle," "accede" and "exceed."

Section III. Twenty matters of form and usage—address, wording of types letterhead, salutation, complimentary close, etc.

Section IV. A test of grammar and usage. The subject must discover such errors in sentences.

Section V. Three short letters which are to test recognition of effectiveness. These are (1) a complaint, (2) a reply to a request for information, and (3) a recommendation. Each sentence is written in three forms: (1) one lively but crude, (2) one affected and wordy, and (3) one direct and sincere. The subject must choose one of the three forms for each sentence. Up to the present there is no reliability reported but the test's length is assurance of its satisfactory reliability. Norms based on 1,200 cases are being improved. Since norms are calculated for both the parts and the total, there is some opportunity for analyzing errors which occur.

In the areas of business arithmetic, business law, and economic geography three tests are simply included in the list.

TESTS OF GENERAL BUSINESS CONTENT

1. General Test of Business Information, grades 9-16. 1942-1943. Forms A and B. Time: 40-45 minutes. Author: Stephen J. Turille. Bureau of Educational Measurements, Kansas State Teachers College Emporia, Kans.

2. Business Fundamentals and General Information Test, United-NOMA Business Entrance Tests, schools and industry. 1939-1947. New Test each year. Time: 45-55 minutes. Authors: Joint Committee on Tests representing United Business Educational Association and NOMA. National Office Management Association, New York.

3. Cooperative Commercial Arithmetic Test, first and second semesters. 1944-1947. Forms U and X. Separate answer sheets. Time: 40-45 minutes. Cooperative Test Service, New York.

4. Examination in Business Arithmetic, high school. 1944. Form B. Separate answer sheets. Time: 135-145 minutes. Authors: Examination Staff of the U.S. Armed Forces Institute. Coopera-

tive Test Service, New York, and Science Research Associates, Chicago.

5. Examination in Business English, high school level, grades 11-12. 1944. Form B. Separate answer sheets. Time: 120-125 minutes. Authors: Examination Staff of the U.S. Armed Forces Institute. Cooperative Test Service, New York, and Science Research Associates, Chicago.

6. Parke Commercial Law Test, high school. 1933. One form. Time: 40-45 minutes. Author: L. A. Parke. Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kans.

7. Primary Business Interests Test, grades 9-15 and adults. 1942. One form. Nontimed. Author: Alfred J. Cardall. Science Research Associates, Chicago.

8. Tate Economic Geography Test, high school level, grades 9-16, 1940. Time: 50-55 minutes. Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kans.

SUMMARY

Objectives in business education are somewhat complicated by the two aims of vocational competence on the one hand and consumer education on the other. The measuring instruments constructed have taken little cognizance of these somewhat conflicting aims. The measuring instruments were divided into (1) tests of clerical aptitude, (2) tests of clerical achievement, and (3) tests of content. Clerical aptitude was measured by tests of discrimination, speed of writing, phonetic spelling, vocabulary, etc. Achievement tests partook of the nature of actual clerical work in an office—taking and transcribing dictation, typing a letter or table, and entering or correcting actual items in a journal or ledger. Tests of content sampled the general information which was needed either to learn business or to understand its general characteristics. In bookkeeping are illustrated both skill and content. An interesting illustration of sound procedure occurs in the United-NOMA Business Entrance Tests. The series of tests bearing this name was composed by a joint committee of the United Business Educational Association and the National Office Management Association. Their tests, given under standard conditions, indicate the proficiency necessary to enter directly into clerical work.

QUESTIONS AND EXERCISES

1. Compare the emphasis of instruction in a consumer-education class with that in a class preparing to enter business.
2. Describe the type of items which are placed in a test of stenographic aptitude. To what uses can an aptitude test be put?
3. How is it possible to validate an achievement test?
4. How do the types of items in an aptitude test differ from those in an achievement test?
5. Why can it be said that bookkeeping involves both skill and content?
6. Explain and illustrate the difference between a test of skill and one of content.
7. What are three characteristics of the tests constructed by the Examination Staff of the U.S. Armed Forces Institute?
8. What are the general purposes of the United-NOMA Business Entrance Series? What characteristic makes them of small use to the classroom teacher?
9. What are other functions of stenographers in addition to taking and transcribing dictation?

BIBLIOGRAPHY

ANDERSON, ROY N.: "Review of Clerical Tests (1929-1942)," *Occupations* (1943) 21:654-660.

BARRETT, DOROTHY M.: "Prediction of Achievement in Type-writing and Stenography in a Liberal Arts College," *Journal of Applied Psychology* (1946) 30:624-630.

BINGHAM, WALTER VAN DYKE: *Aptitudes and Aptitude Testing*, Chaps. XII, XIII, pp. 322-329. New York, Harper & Brothers, 1937.

BLACKSTONE, E. G.: "Commercial Education," *Encyclopedia of Educational Research*, pp. 426-440. New York: The Macmillan Company, 1941.

BUROS, OSCAR K. (ed.): *The Third Mental Measurements Yearbook*, Items 365-396, 623-632. New Brunswick, N.J.: Rutgers University Press, 1949.

———: *The Nineteen Forty Mental Measurements Yearbook*, Items 1476-1491, 1664-1665. Highland Park, N.J.: The Mental Measurements Yearbook, 1941.

———: *The 1938 Mental Measurements Yearbook*, Items 935-945. New Brunswick, N.J.: Rutgers University Press, 1938.

GREENE, HARRY A., ALBERT N. JORGENSEN, and J. RAYMOND GERBERICH: *Measurement and Evaluation in the Secondary School*, Chap. XXII. New

York: Longmans, Green & Co., Inc., 1943.

HESLER, RUSSELL J.: "Aptitude Testing in Shorthand," *Journal of Business Education* (1947) 22:25.

JURGENSEN, CLIFFORD E.: "A Test for Selecting and Training Industrial Typists," *Educational and Psychological Measurement* (1942) 2:409-425.

KLUGMAN, SAMUEL F.: "Test Scores for Clerical Aptitude and Interests before and after a Year of Schooling," *Journal of Genetic Psychology* (1944) 65:89-96.

MORROW, ROBERT S.: "An Experimental Analysis of the Theory of Independent Abilities," *Journal of Educational Psychology* (1941) 32:495-512.

SCHNEIDLER, GWENDOLEN G.: "Grade and Age Norms for the Minnesota Vocational Test for Clerical Workers," *Educational and Psychological Measurement* (1941) 1:143-156.

——— and DONALD G. PATERSON: "Sex Differences in Clerical Aptitude," *Journal of Educational Psychology* (1942) 33:303-309.

TONNE, HERBERT A.: *Consumer Education in the Schools*, Chap. 8. New York: Prentice-Hall, Inc., 1941.

TURSE, PAUL L.: "Problems in Shorthand Prognosis," *Journal of Business Education* (1938) 13:17-18.

CHAPTER 12

Measurement of Fine Arts and Manual Arts

These two areas of fine arts and manual arts are grouped together in part for convenience and in part because there is a certain affinity between them. Performance in music and art is directly related to manual facility, while much of the success in manual arts is due to the artistic manner in which the object is constructed.

In this chapter, we shall consider the measurement and evaluation of (1) music, (2) art, and (3) manual and mechanical arts and home economics.

MUSIC

The world of music is practically universal. What was in the past a rather select affair where people foregathered in concert hall, opera, or academy of music has now become ubiquitous. Bands in schools, and at games of various kinds; the movies, and perhaps above all the radio and television have bombarded us with music of some kind daily and continuously. Music has come to occupy the largest place in our leisure-time activities. Under these conditions the school has no other course than that of introducing its charges to this world of music.

There are two major aspects of measurement concerned with music: (1) the measurement of aptitude or talent, and (2) the measurement of achievement. A third aspect, that of appreciation, is not coordinate with the first two but, for the general population, may be of equal if not superior importance.

MEASUREMENT OF TALENT IN OR APTITUDE FOR MUSIC

The measurement of musical talent takes its beginning from the experimental work of Carl Emil Seashore who, after years of experimentation, published his results in 1919 under the title *The Psychology of Musical Talent*. In this book he sets forth both an analysis of musical talent and a description of the procedures for measuring it. The musical mind is made up, in part, said he, of (1) the sense of pitch, (2) the sense of intensity, (3) the sense of time, (4) the sense of rhythm, (5) the sense of consonance, and (6) tonal memory. He first demonstrated how these traits were measured by tuning forks and complicated laboratory

apparatus and second, and perhaps more importantly, described the phonograph records on which with minute exactness were impressed the same procedures.

In 1939 a revision of the tests was published in Seashore's *Measures of Musical Talents*, revised edition. These new tests embodied the main features of the original test changing only the test of consonance to one of timbre. The revised edition calls these divisions pitch, loudness, time, tonal memory, timbre, and rhythm. It also has two series, A and B. Series A is intended to test the capacities of unselected groups of children or adults. Series B measures the capacities of more specialized groups such as musicians and prospective musicians. The test for each series is furnished on three 12-inch phonograph records with a complete test on each side of the record.

What of this test of musical talents? Is it reliable? Does it really measure musical talents? The directions are clear: "You will hear two tones which differ in pitch. You are to judge whether the second is higher or lower than the first. If the second is higher, record H; if lower, record L." It is generally added "If you are not sure, guess." The *reliability* of these measures is indicated by coefficients of correlation which for the individual tests vary from .62 to .89. The coefficients are highest for tonal memory, pitch, and loudness. The constructor recommends that these six scores not be combined into one total score but that each one be treated as a separate entity in the formation of a profile of musical talent. Norms are provided for grades 5 to 8 and for adults. If we apply our strictest principles to these measures of reliability we see that they are not reliable enough to discriminate between the aptitudes present in the same individual. For such a purpose a coefficient of .90 to .95 is required. Another even more fundamental question is whether these measures, gathered in a more or less artificial manner, operate in music as they do under the testing conditions.

The answer to this last question is indicated by the correlations and uses which are now introduced and together constitute some measure of the tests' validity. About the only criterion available against which to measure tests of musical talent is success in courses in music. These may be the more theoretical courses in harmony or counterpoint or the more practical courses dealing with instruments. Success in such courses is determined by factors of interest, ambition, intelligence, and previous training as well as by fundamental musical talent. For this reason, the correlation coefficients between measures of musical talent and success in courses in music has not been very high. In reviewing 16 studies which had been completed up to that time (1931), Farnsworth¹ reports the

¹ Farnsworth, P. R., "An Historical, Critical and Experimental Study of the Seashore Kwalwasser Test Battery," *Genetic Psychology Monograph* (1931) 9:291-389.

correlations with school marks in music as varying from $-.08$ to $.45$. When each of the measured traits of the Seashore tests is correlated with school marks after one semester the following correlations resulted.¹

| | |
|-------------------|------|
| Pitch..... | .11 |
| Intensity..... | .07 |
| Time..... | .20 |
| Consonance..... | -.27 |
| Tonal memory..... | -.19 |
| Rhythm..... | .25 |

In general, these results are much lower than the usual correlations found. The trend can be more readily inferred from Table 9. An inspection

TABLE 9. PREDICTING SUCCESS IN THE STUDY OF MUSIC*

| | Achievement in musical theory | Median r | Sight singing, ear training, or dictation | Median r | Achievement in applied music | Median r |
|---------------------------|-------------------------------|------------|---|------------|------------------------------|------------|
| Time..... | .13-.56 | .29 | .02-.56 | .29 | .10-.63 | .23 |
| Pitch..... | .03-.64 | .38 | .03-.64 | .54 | .01-.62 | .18 |
| Tonal memory..... | .16-.70 | .36 | .23-.70 | .57 | .01-.65 | .19 |
| Intensity..... | .05-.40 | .30 | .05-.40 | .30 | .07-.50 | .06 |
| Rhythm..... | .14-.39 | .21 | .14-.39 | .21 | .06-.52 | .20 |
| Consonance..... | .05-.37 | .28 | .05-.37 | .29 | -.27-.52 | .06 |
| Total scores, Seashore... | .21-.75 | .44 | .40-.70 | .46 | -.15-.31 | .13 |
| Mental-ability tests..... | .23-.66 | .41 | .23-.64 | .29 | .03-.32 | .33 |

* *Predicting Success in the Study of Music*, Veterans Administration Technical Bulletin TB 7-77, Dec. 21, 1947.

tion of this table is very revealing. Let us take first the median correlations. The median correlations between school marks in musical theory and the Seashore Measures of Musical Talent range from .21(rhythm) to .36(tonal memory) and .38(pitch). You will note that the Combined scores, though not recommended by Seashore, give the highest coefficient, .44. Note also that an intelligence test is as good for predicting success in musical theory as are the tests of musical talent. When we turn to sight singing, ear training, and dictation the tests are more efficient. The median coefficients in this instance range from .21 to .57. It is evident that pitch and tonal memory taken individually stand out

¹ Mursell, James L., *The Psychology of Music*. New York: W. W. Norton & Company, 1937.

clearly above the others and even above a combination of the six in a total score. Intelligence tests, too, are far below the talent tests in the area of sight singing and ear training. When we consider applied music no one of the individual tests or their combination furnish any real aid in prediction. The tests of intensity and consonance have no more than chance, or zero, correlations with marks in applied music. Time, the highest, with a coefficient of .23 shows only a low relationship. Mental-ability tests with a coefficient of .33 are distinctly more closely related to success in the area of practical music than are Seashore's tests.

From the previous discussion, it may be inferred that for predicting success in music some combination of intelligence tests and musical tests might be better than either alone. We are fortunate in having an extended investigation of the combination of the Iowa Comprehension Reading Test and the Seashore tests in predicting musical success in a standard musical college.¹ In this study it was established through preliminary investigations that it was practical to divide the entering students into five groups—(1) safe, (2) probable, (3) possible, (4) doubtful, and (5) discouraged—on the basis of their standings in the two tests (Seashore's Measures of Musical Talent and Iowa Test of Silent Reading). If the students were very low on both tests they were to be discouraged in their intention to proceed with their musical education. If they scored very high on both, then they were safe as far as their prospects for success and graduation went. The following data show the probability of graduation achieved by each group:

| Group | N | Percent of graduated |
|------------------|-----|----------------------|
| Safe..... | 125 | 60 |
| Probable..... | 143 | 42 |
| Possible..... | 195 | 33 |
| Doubtful..... | 73 | 23 |
| Discouraged..... | 29 | 17 |

Furthermore the students with high scores stayed in school longer, had fewer dismissals, gathered in more of the honors, and made more recital appearances than did those who received low scores. It seemed clear that this combination of intelligence test and musical-talent test was a practical success for selecting students for advanced musical training.

Other combinations which include the Seashore test show considerable efficiency in prediction. In one study a combination of Seashore's tests, Henmon-Nelson Intelligence Test, and Teachers College Achievement

¹ Stanton, Hazel Martha, *Measurement of Musical Talent*, Studies in the Psychology of Music, Vol. II, University of Iowa, 1935.

Test correlated .84 with school marks in sight singing received by college students.¹ But it must also be mentioned that weighted scores on Seashore's pitch and tonal memory correlated .72 with marks in sight singing in the case of 131 students, while a combination of Thurstone's Intelligence Test, Iowa High School Content Examination, and Seashore's pitch and tonal memory tests correlated .43 with marks in the history and appreciation of music. From an inspection of the results of these combinations one can conclude that the right combinations of intelligence-test scores and musical-test scores are highly successful in predicting success in certain aspects of musical training.

In concluding about the efficiency of the Seashore Measures of Musical Talents in comparison with other like measures, the following quotation is approximately correct:²

The battery is so much better in almost every way than its chief rivals, the Tilson-Gretsch and the Kwalwasser-Dykema, that music testers should use it exclusively in their attempts to screen out those unfortunates who will not achieve success in music without enormous effort.

A second test of musical talent, the Kwalwasser-Dykema Music Tests for grades 4 to 16, like the Seashore tests are imprinted on phonograph records. The present test uses five double-disk records, by means of which the following tests may be given:

1. Tonal memory
2. Quality discrimination
3. Intensity discrimination
4. Feeling for tonal movement
5. Time discrimination
6. Rhythm discrimination
7. Pitch discrimination
8. Melodic taste
9. Pitch imagery
10. Rhythm imagery

An inspection of the list of tests shows that seven of the tests cover the same areas as does the Seashore Measures of Musical Talents, but the last three are new.

These tests are claimed to be "indicative of musical talent and achievement." In the manual (1930) norms are furnished but no data

¹ See *Predicting Success in the Study of Music*, Veterans Administration Technical Bulletin TB7-77, Dec. 21, 1947, in which there are summaries of many studies of combinations.

² Farnsworth, Paul R., in a review in *The Third Mental Measurements Yearbook*, p. 177. New Brunswick, N.J.: Rutgers University Press, 1949.

are presented on validity or reliability. The test is easily administered and scored. It has been rather widely used by music educators.

A new test of music has recently (1950) appeared, Musical Aptitude Test, by Harvey S. Whistler and Louis P. Thorpe.¹ The constructors of this test discard the analytic approach of Seashore and declare that "rhythm, pitch, and melody are the basic elements of all music."² "The test is divided into five parts for administration: (1) rhythm recognition,

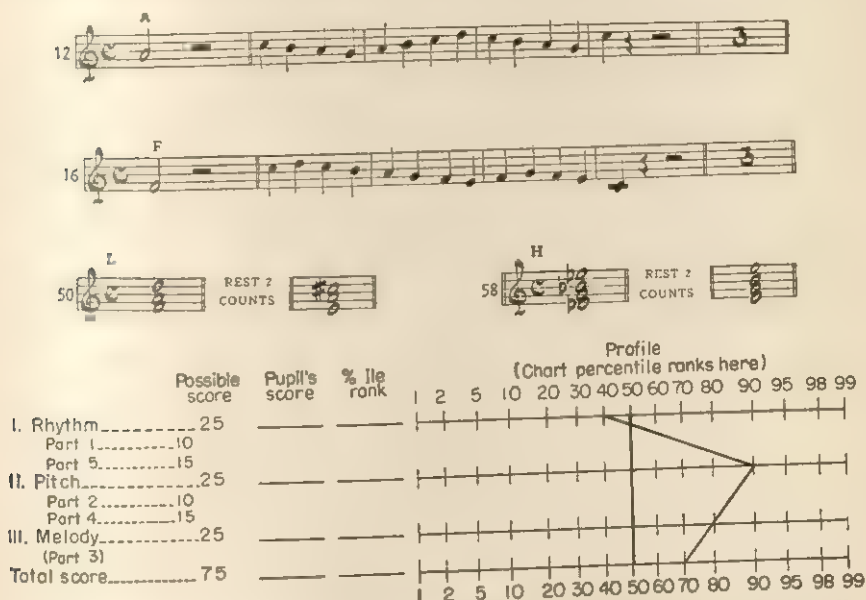


FIG. 18. Musical Aptitude Test: pitch recognition, pitch discrimination, and profile. (Whistler and Thorpe, 1950.)

10 items; (2) pitch recognition, 10 items; (3) melody recognition, 25 items; (4) pitch discrimination, 15 items; and (5) advanced rhythm recognition, 15 items." All tests are played upon the piano and are responded to on a separate sheet. The time needed for taking the test is about 50 minutes. To pairs of melody samples and to pairs of rhythm samples the subject responds S (same) or D (different). To the test of pitch recognition the subject responds 0, 1, 2, 3, 4, according to the number of times a tone occurs in the melody. (Consider the two samples from the test of pitch recognition shown in Fig. 18.)

In the test of pitch discrimination the subject responds to the two chords presented with a two-count rest between, with S (same), H (high),

¹ California Test Bureau, Los Angeles, Calif. Items by permission.

² Quotations from the manual.

or L (low). From the scores, a profile may be drawn, as shown in Fig. 18, on the furnished chart.

There are reported in the manual the results of the studies performed in standardizing the test. The *validity* was studied by correlating the total test scores with teachers' estimates of instrumental talent ($r = .37$) and of vocal talent (.56), and with whether subjects had played on an instrument for 1 year ($r = .56$) in an orchestra or band ($r = .40$) or had sung in a chorus, choir, or glee club ($r = .19$). The last three correlations mean that there was a tendency (as indicated by the size of the coefficient) for subjects who had played an instrument for 1 year or more to make high scores, for example, and for those who had not played an instrument that long to make low scores. The *reliability* for the total score is reported as .93 and for the three divisions as .80 to .88. Percentile norms have been calculated from 2,000 cases. The "data were corrected so that the average I.Q. of the standardization population was 100 and the standard deviation of the distribution of the I.Q.'s was 16 for grades 4-8 inclusive."¹ The test correlated only .156 with intelligence when the chronological age was made relatively constant.

Here, then, is a new test of musical aptitude described in greater detail because of its newness. It has not yet been studied adequately. Until a test has been correlated with many variables no one *can know* whether it will prove useful or not.

Tests of musical aptitude aid the teacher and counselor (1) to advise with students concerning their study of instrumental and vocal music, (2) to aid in the grouping of students for purposes of instruction, and (3) to advise with students about pursuing a musical career.

INFORMATION, APPRECIATION, AND ACHIEVEMENT

It must be remembered that a successful achievement test indicates the amount of progress achieved by a student toward a defined objective. The objectives in the teaching of music were clearly defined and printed in 1921.² The following summary of the eighth-grade attainments may give us later an idea of how well these objectives have been worked out. The educational council of music supervisors not only laid down the attainable objectives but mentioned the number of students who could reasonably be expected to attain them. It must be recognized that a smaller number of individuals can be brought to sing songs alone than can be taught to sing them in a group.

¹ *Manual*.

² *Report of Educational Council of the Music Supervisors' National Conference*. Washington, D.C.: National Education Association, 1921.

Attainable Objectives for Grade 8

1. Ability to sing well and with enjoyment 30 to 50 (a) unison, (b) two part, and (c) three-part songs. This group includes community and national songs. About 90 per cent of individuals sing alone at least 10 of these songs.

2. Ability to sing at sight, using words, a unison song of hymn-time grade; or, using syllables, a two-part song of hymn-time grade and easiest three-part songs. About 30 per cent of pupils sing these songs individually.

3. Ability to appreciate the charm of design of songs sung; to give the salient features of structure in a standard composition; to identify a three-part song after hearing it a few times and to know the titles and composers of 20 standard compositions.

4. Knowledge of essential facts of elementary theory so that 75 per cent of students can give correct explanation of notational features in pieces of average difficulty.

One may say more briefly that the attainable outcomes of instruction in music may be thought of as moderate amounts of success in (1) singing well, (2) singing at sight, (3) appreciating the charm and design of songs, and (4) acquiring enough knowledge of theory to give correct explanations of notational features. There have been attempts to measure outcomes in each of these areas. For measuring the ability to sing well there is the Mosher Test of Individual Singing.¹ In this test 12 exercises arranged in order of difficulty are to be sung by the subject and scored by judges according to definite instructions. There is also the Hillbrand Sightsinging Test.² This test for grades 4 to 6 contains six songs in a four-page folder. The pupil studies the songs for a few minutes, and then sings them without help or accompaniment. There are nine different kinds of errors which, when made, are to be recorded on a copy of the songs. The errors are:

1. Notes wrongly pitched
2. Transpositions
3. Times flatted
4. Times sharpened
5. Notes omitted
6. Errors in time
7. Extra notes

¹ Mosher, Raymond M., *A Study of Group of Measurement of Sight-singing*, Contributions to Education, No. 194. New York: Bureau of Publications, Teachers College, Columbia University. 1925.

² Hillbrand, E. K., *Hillbrand Sightsinging Test*. Yonkers, N.Y.: World Book Company, 1923.

8. Repetitions

9. Hesitations

Probably the most complete test for the knowledge of school music is the Kwalwasser-Ruch Test of Musical Accomplishment.¹ It is intended for grades 4 to 12. In its construction, attempt was made to use items from representative courses of study. There are 10 divisions of the test:

1. Knowledge of musical symbols and terms. To answer the items requires a knowledge of the tones of the scale, flats, sharps, clefs, rests, crescendo, diminuendo, *lento* and *legato*. For example,

19. *Allegro* means lively slow repeat accent sweetly

2. Recognition of syllable names from notation. A variety of staves with six different notes for staff. "Write the *syllable names* on the lines under the other notes."

3. Detection of pitch errors in the notation of a familiar melody. Five wrong measures to be detected in two lines of notes.

4. Recognition of time errors in the notation of a familiar melody. Detection of five measures that have the wrong number of beats in two lines of notes of the song "America."

5. Knowledge of pitch or letter names of bass and treble clef. Two lines of notes in the treble clef; two in the bass clef. There are five notes to the line for which the subject must write the pitch or letter name.

6. Knowledge of time signatures. Test to discover the time signatures for each of 10 full measures.

7. Knowledge of key signatures. Must write the names of each of 10 major and 5 minor key signatures.

8. Knowledge of note values. Draw a line under one of five notes which is needed to complete each of five measures.

9. Knowledge of rest values. Subject draws a line under the rest needed to complete each of five measures.

10. Recognition of familiar melodies from notation. Subject writes out the name of the song from each of ten lines of notes.

The reliability of this test was .97 for 167 children from the sixth, eighth, tenth, and twelfth grades. This probably would be reduced to .92 or .93 if the children had all been selected from one grade. Available norms are based on some 5,000 children.

This test covers *only* the informational and factual sides of the objectives set up by the Music Supervisors' National Conference of 1921. Such acquisition of facts is related to intelligence almost as much as to musical ability. *A person who scores high on this test would not necessarily stand high in musical accomplishment.*

¹ Published by the Extension Division of the State University of Iowa, 1924 and 1927.

TESTS OF INFORMATION AND APPRECIATION

An indication of interest in music can be had from one of the divisions of the Kuder Preference Record. It is also somewhat indicated by an acquaintance with the authors of great music as well as with the music itself. For this reason Kwalwasser's Test of Music Information and Appreciation¹ is interesting. This test is divided into three major divisions: (1) history and biography, (2) instrumentation, and (3) musical form.

Under History and Biography there are tests involving the classification of such artists as Galli-Curci, Louis Graveure, Albert Spalding, Hans Kindler, and John Powell under (1) vocalists, (2) pianists, (3) violinists, (4) cellists, and (5) conductors. Another test inquires about the nationality of composers, while another asks who were the composers of famous compositions. The final test in this division consists of 50 true-false items based on the general knowledge of composers and compositions. Illustrations are:

6. Liszt expanded the range of pianism
16. Mozart became deaf during the last years of his life
23. The metronome is associated with the name of Maelzel
33. Chopin wrote exclusively for the voice
41. The symphonic poem was originated by Bach

Division II, on Instrumentation, asks whether tones on 10 orchestral instruments are produced by (1) blowing, (2) striking with hammers, or (3) bowing, *e.g.*, oboe, viola, bassoon, melaphone. The subject is also asked to classify 10 orchestral instruments into (1) string section, (2) wood-wind section, (3) brass-wind section, and (4) percussion section. Such instruments as violin, xylophone, bassoon, celesta, and ophicleide are mentioned. There are also 50 true-false items which test information. Illustrations are:

1. Viola is an alto horn
10. The bassoon has a double reed
14. The clarinet employs a single reed
24. The euphonium has two "bells" or "flares"
34. Mutes are used only with stringed instruments
44. The bass-viol is usually employed in string quartets

The third section contains 50 true-false items on musical form. Examples are:

¹ Items by permission of Bureau of Educational Research and Service, University of Iowa.

4. An *overture* is played at the end of the opera
14. *Arias* are found in symphonies
24. *Arpeggio* means a gradual increase in loudness
34. The *cantata* is a choral work with solos eliminated
44. The *concerto* is built on the *rondo* form

There are two criticisms of the use of such a test for the measurement of appreciation. The first, a minor one, finds in the test constructed several years ago a lack of modernity in some of the items, *e.g.*, the classification of Galli-Curci as an artist. The second criticism questions whether a test of general information about music is really an indication of appreciation. Information is notably correlated with intelligence as will be shown in Army Alpha and in the more recent test, Wechsler-Bellevue. There is no doubt that intellectual capacity does enter into the scores of the test which has just been described. Just what part of the test is appreciation and what part intelligence has not been determined.

ART

The enjoyment of beauty is as old as civilization itself. For many years art was thought of as connected with the greatest productions of mankind such as the temples along the Nile, the Parthenon in Athens, or Da Vinci's "Last Supper." In recent years, two great movements have favored a more universal application of the principles of art. The first of these is the realization that beauty of form and color can apply to the great majority of objects with which we are surrounded. It was more and more evident that the surroundings of even small houses could be beautiful and that arrangement of mass and color could be carried out with trees and shrubs and flowers. The house itself could be made beautiful both as to its exterior and interior. Clothes, utensils, public buildings, stores, even garages could also help to furnish a beautiful appearance to a town. In hundreds of other areas, too, beauty could be both present and appreciated. The second development that favored a greater interest and appreciation of art was the discovery that children could express their imitative capacities as well as their creative imaginations in art forms. Sometimes in a very crude drawing an idea thus might take shape and when supplemented by verbal explanations partake of the nature of art. Truly, we have been late in realizing with Keats that "a thing of beauty is a joy forever" and that its loveliness increases.

OBJECTIVES IN THE TEACHING OF ART

The objectives of the teaching of art may be roughly summarized as follows:¹

¹ See Whitford, W. G., *An Introduction to Art Education*. New York: Appleton-Century-Crofts, Inc., 1929.

1. To acquire the knowledge of the principles of art and of their application to everyday experiences: (a) In fine arts, to attain the knowledge of the principles used in the construction of great pictures, architecture, sculpture, etc.; and (b) in applied arts, to learn the principles of art as used in the construction and making of furniture, clothing, interior decoration, dishes, utensils, etc. In brief, this means the application of the knowledge of line, mass, and color to the everyday experiences of life. This results in good taste and discriminating judgment when choosing objects.

2. To secure an appreciation of the beautiful wherever found: (a) In flowers, sky, ocean, trees, buildings, clothes, birds in flight, painting, buildings, and in modern products of all kinds; (b) in the various attempts to add beauty to a community through community centers, store fronts, art galleries, etc.; and (c) in various attempts to beautify both the interiors and exteriors of homes.

3. To get some experience in and capacity for creating beauty: (a) in selecting and grouping fine objects for specific purposes and in securing some originality in the process, and (b) in acquiring some skill in drawing and painting objects which conform to art and verity. This involves the coordination of eye, hand, and idea.

4. To develop keener capacities for observation so as to discover beauty in nature. Knowledge of what to look for and how to judge beauty or its absence in the objects which surround us. The teacher must stimulate whatever capacity the child possesses in the way of originality, initiative, and imagination in dealing with objects around him.

It will be seen that hardly any test covers adequately more than a sizable fraction of these objectives.

MEASUREMENT IN ART

Measurement in art, as in music, has two aspects: (1) the measurement of capacity, and (2) the measurement of achievement.

The Measurement of Capacity

In attempting to measure the capacity of subjects for the learning of art, test constructors have tried to analyze the total product into a few fundamental processes, which, if they are done well, indicate probable success in this undertaking. Three measures of capacity are described here: (1) the Meier-Seashore Art Judgment Test (125 pairs of pictures) and its revision, the Meier Art-Judgment Test (100 pairs of pictures with new scoring), (2) the McAdory Art Test, and (3) the Lewerenz Tests in Fundamental Abilities of Visual Art.

The Meier-Seashore Art Judgment Test and its revisions by Meier grew out of six years of experimentation and subsequent revision. The

125 pairs of items are the survivals of some 600 drawings after critical tryouts and judgments by experts. The art forms of the pictures have stood the test of time for they were adapted from the work of old masters, from contemporary artists, and from Japanese prints. According to the manual, all items (1) were from reputable works, (2) exemplified aesthetic principles, and (3) were suitable for testing purposes. In taking the test, the subject, with the name of the picture and two pictures before him, indicates his preference by drawing a circle about the L if



FIG. 19. Pictures used to indicate preference for drawings, Meier-Seashore Art Judgment Test. (By permission of Bureau of Education Research and Service, University of Iowa, Iowa City.)

he decides the left-hand picture is better, or around the R if he believes the right-hand picture more desirable. You will note that there is only one thing different in the pair. One member of each pair is as the artist drew it (Fig. 19). The judgment is made on 125 pairs.

The reliability, validity, and norms of the test are well worked out. The coefficients of reliability range from .71 to .85 when the test is repeated. Its validity has been carefully studied. Differences appear in test scores where there are differences in art achievement. For example, the authors report a median of 87 for the art faculty, 82 for art students, 76 for the twelfth grade; 72 for the tenth; and 66 for the eighth. Some of the children in Grades 8 and 10 scored as high as the experts. Whatever is measured by this test correlates very low with intelligence. The correlations with intelligence tests vary from $-.14$ to $.28$ with a median coefficient of $.16$ or $.17$. Substantial correlations however, have been found with marks in art classes at the college level. Percentile norms are

available for Grades 7 and 8, 9 and 10, and 11 and 12. It is also asserted that those who score in the highest quarter (percentiles 76 to 100) are almost certain of success; those in percentiles 51 to 75 will profit from instruction and have a chance at an art career, those in percentiles 26 to 50 may be able to do the manual part of drawing, and those below the 25th percentile should retake the test. These last will probably not succeed in art.¹

The authors believe that this test measures aesthetic judgment, the most essential characteristic of artistic production. "Aesthetic judgment is defined as the capacity for perceiving quality in aesthetic situations relatively apart from formal training." The items are rather permanent in nature so that time does not affect them greatly. One critic² believes the test measures the perception of quality rather than its production. It "represents a useful measure of individual sensitivity to aesthetic organization of graphic form." Some critics wonder if a test constructed almost entirely of the graphic arts can apply to the whole field of art and whether there are not other factors in artistic competence that are just as important.

The McAdory Art Test is another instrument for measuring art judgment. It differs from the Meier-Seashore Art Judgment in several particulars. In the first place the materials out of which the test is constructed are of a practical nature, made up of samples of texture and clothing, architecture, furniture and utensils, as well as of dark and light masses, paintings, and shape and line arrangements (Fig. 20). In each plate, there are four samples—A, B, C, and D—which the subject is to arrange in the most pleasing order. He receives one point for each sample which he judges to be in the position voted by expert judges. The whole test has been restudied and the samples judged again by 30 art experts.³ In this revision, four plates were eliminated and the positions were changed in four others. All told there are 72 plates, 24 of which are in color. By means of record sheets on which the judgments can be registered this test may be given to as many as 30 students at one sitting.

The reliability of the test varies from .79 to .93 depending on the population which is used. Its validity has been studied by relating its scores to other art tests. For example its correlation with the Christensen

¹ See *Examiner's Manual*, pp. 8-9, 1930.

² Saunders, A. W., *Third Mental Measurement Yearbook*, *op. cit.*, Item 1327.

³ See Siceloff, Margaret McAdory, and Ella Woodyard, *Validity and Standardization of the McAdory Art Test*. New York: Bureau of Publications, Teachers College, Columbia University, 1933; and McAdory, Margaret, *The Construction and Validation of an Art Test*. New York: Bureau of Publications, Teachers College, Columbia University, 1929.



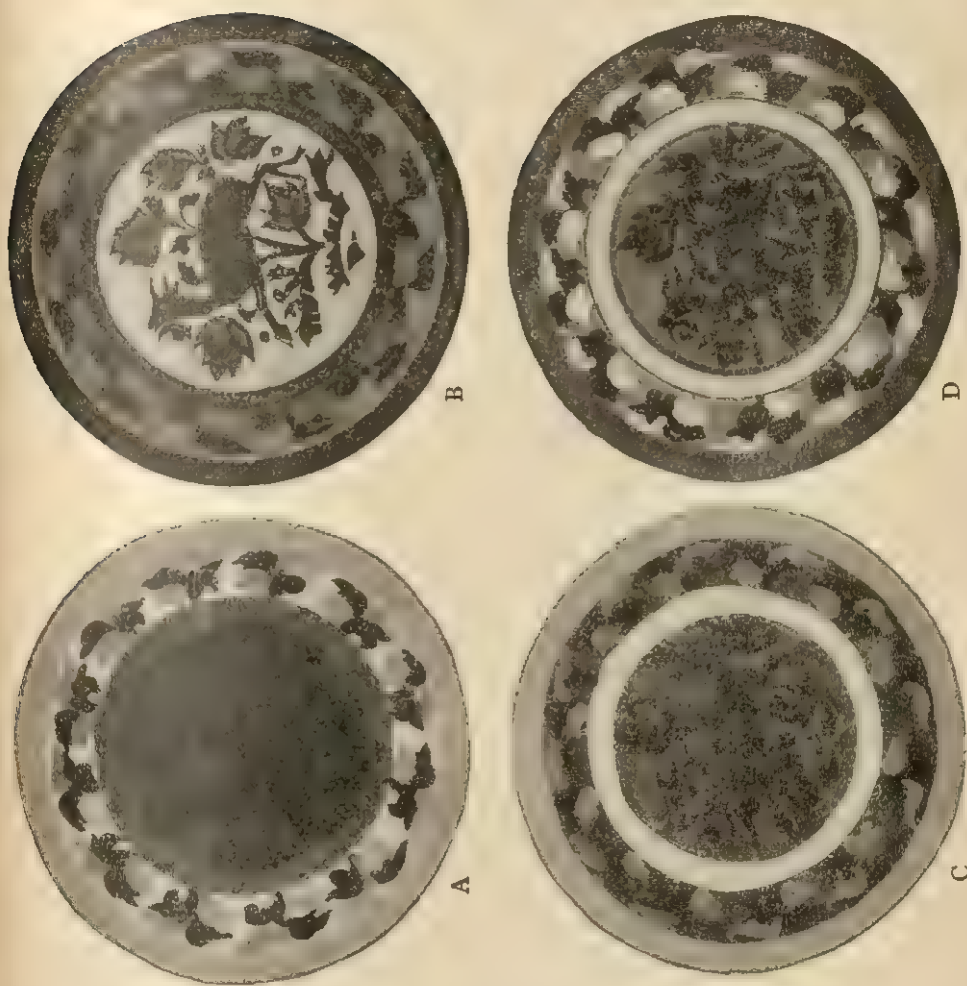


FIG. 20. Plates 8 and 19 of the McAdory Art Test. (By permission of Margaret McAdory Siceloff.)

Art Test was .63, with the Meier-Seashore Art Judgment Test, .27, and with the Levering Art Judgment Test, .58. The author explains the low correlation with the Meier-Seashore test by saying that art appreciation is dependent upon the particular objects judged. As far as the author knows, few, if any, correlations have been computed with achievement in artistic occupations. Norms have been established from the measurement of 5,000 or 6,000 students in the New York area and extend from grade 3 to college and art schools. As with other art tests, its correlation with *intelligence* is low (.15).

According to the author, the uses of the tests are varied. As an educational device it distinguishes those with artistic ability from others who do not possess it. It can thus select pupils for art classes as well as help the teacher decide whether art work should be continued. It may be used when advising with students concerning their prospective occupations which require ability in art. In the third place, it may have consumer use in helping the ordinary individual to know how much dependence to put on his own judgment in selecting art objects for daily use.

The value of this test is lessened because styles are continually changing in the practical materials of which its plates are composed. It has the advantage of being a group test. Its correlations, however, with college teachers' ratings of art students are below that of the Meier-Seashore test and are low with other art and intelligence tests. In general, then, the McAdory Art Test for ordinary purposes would rank below the Meier-Seashore Art Judgment Test.

The Lewerenz Tests in Fundamental Abilities of Visual Arts,¹ grades 3-12, are divided into three parts, as shown in the accompanying table.

| | Part | Time, minutes |
|-----|---|---------------|
| I | | |
| | 1. Recognition of proportion..... | 10 |
| | 2. Originality of line drawing | 20 |
| II | | |
| | 3. Observation of light and shade..... | 5 |
| | 4. Knowledge of subject-matter vocabulary..... | 20 |
| | 5. Visual memory of proportion..... | 5 |
| III | | |
| | 6. Analysis of problems in cylindrical perspective..... | 5 |
| | 7. Analysis of problems in parallel perspective..... | 5 |
| | 8. Analysis of problems in angular perspective..... | 5 |
| | 9. Recognition of color..... | 20 |

In Test 1 the subject selects from the same object represented in four

¹Lewerenz, Alfred S., *Lewerenz Tests in Fundamental Abilities of Visual Arts*, California Test Bureau, Los Angeles, Calif. Items by permission

different proportions that one which is the best. There are cups, friezes, cornices, curves, masses, etc., each with four proportions from which the subject must select the best. Test 2 consists of 10 sets of dots arranged in a haphazard manner through which the subject is to draw interesting things. This is a fine test of originality in imagination. In Test 3 (Fig. 21) the subject marks with an X those areas where there should be shade. Such objects as cubes, spheres, cylinder and cup, and a house are the

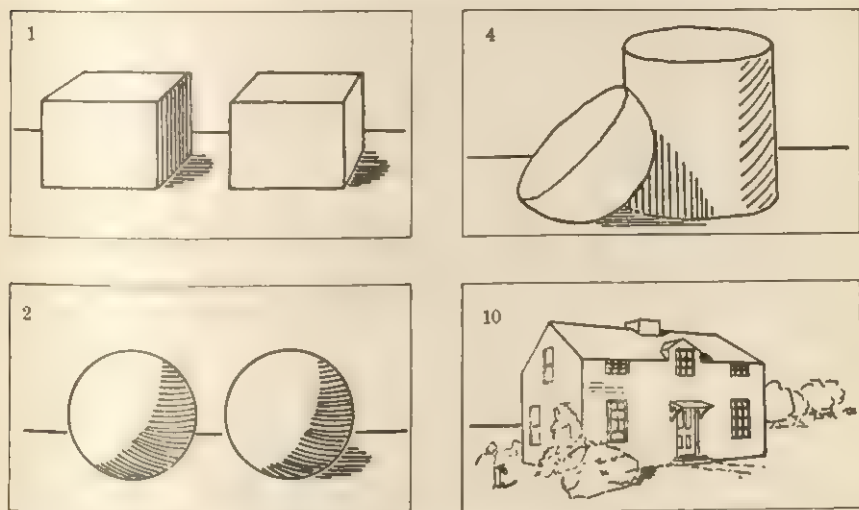


FIG. 21. Lewerenz Tests in Fundamental Abilities of Visual Arts. Observation of light and shade.

pictures. The directions, to be read aloud by the examiner and silently by the pupils, are as follows:

This is a test to show how well you understand and interpret problems in light and shade. In the ten drawings below mark with an (X) each place or surface where you think there should be a shade or a shadow. The light is coming from the left. Only the objects in No. 6 and No. 7 are open.

Test 4 is a test of the vocabulary of materials, processes, drawing and of the authors of pictures. Test 5 is a picture of a large vase which after being seen must have its outline drawn. Tests 6, 7, and 8 have to do with different types of perspective. Test 9 is a test of color recognition. Along the top of the color chart are six standard colors: red, orange, yellow, green, blue, and violet. The subject looks at the color on the chart and writes down the standard colors from which the mixed color is formed.

The reliability of the test is indicated by a coefficient of .87 computed on 100 pupils in grades 3 to 9, only a fair figure for reliability. Its

validity has been studied by correlating its scores with semester grades in art. In one case this figure was .40 (manual). In another study, as reported in the manual the coefficient based on test ranking and teacher estimate was .63. Norms are presented for groups of grades: (1) grades 3 to 6, (2) grades 7 to 9, (3) grades 10 to 12, and (4) first-year university art students. Scores are presented for each test at five different levels of attainment which may be translated into verbal descriptions: (1) very superior, (2) superior, (3) average, (4) inferior, and (5) very inferior. On the front sheet of the tests these latter may be graphed to indicate the relationship between the nine tests and different degrees of success in each.

The usefulness of this test will depend on whether the approach to art should come through the accomplishment of a series of separate skills or from the study of integrated wholes such as industrial arts, architecture, or painting. Another problem arises as to whether measures of perspective are really art or merely tools of art. Reviewers agree that the value of this test depends on the philosophy of the teacher who contemplates its use.

Measurement of Achievement

Only one test will be used to illustrate the measurement of achievement.

The Knauber Art Ability Test,¹ measures both capacity and achievement. While many of the tests of art thus far described have consisted of judging or, at most, finishing a drawing the present test consists largely of actual drawing either from memory or imagination. Consider the problems of the test. After the first test, which consists of drawing from memory a rather elaborate design, the major problems are concerned with making original drawings. The subject must draw the figure of Santa Claus; draw a cup in a saucer; arrange a composition of three trees, a cottage, and a path; and draw "The Homeless Dog." These drawings are graded both for composition and for expression of emotion. The author furnishes scales at three levels of quality -10, 6, and 3—by means of which the drawings can be more accurately rated. The reliability of the test is reported as .95 in the case of 83 subjects who varied greatly in ability. The test's validity has been studied. The average score on the test for art teachers was 123, for non-art teachers, 61. The median for art majors in the junior class in college was 95, for non-art majors, 52.² Norms were computed on the basis of grades. For each grade, from the seventh through the twelfth, medians are furnished along with degrees of ability as shown in the accompanying table.

¹ Knauber, Alma Jordan, *The Knauber Art Ability Test*. Cincinnati, Ohio: published by the author.

² From the *manual*.

| Grade | Grade norm | Very low ability | Low ability | Average ability | Ability | Exceptional art ability |
|-------|------------|------------------|-------------|-----------------|---------|-------------------------|
| 7 | 20 | 0-10 | 11-15 | 16-28 | 29-39 | 40-170 |
| 12 | 58 | 0-18 | 29-42 | 43-69 | 70-89 | 90-170 |

Similar records are furnished at the college level. The author claims that this test measures largely native ability. On the surface it is a measure of competency gained in taking courses in art. The scores undoubtedly reflect partly native ability, partly interest, and partly the adequacy of training received.

MANUAL ARTS

More than 40 per cent of the citizens of the United States who are gainfully employed are working directly or indirectly in activities that demand some facility with or knowledge of machines. Thirty per cent of our population are skilled workmen, many of whom need to understand mechanical processes and to manipulate parts of machines. Add to these skilled individuals a goodly number of machine tenders who are semi-skilled. Furthermore, there are a rich variety of occupations in this area, varying in complexity all the way from changing spools on a machine to building a cabinet. For these reasons, the measurement and prediction of mechanical ability or aptitude is of the greatest importance. In the third place, there is an inclination in some quarters to direct students who fail in academic subjects into the courses in manual arts without regard to their mechanical aptitude or ability. While there is only a small correlation between academic aptitude and mechanical aptitude, there is ample evidence to show that those low in academic aptitude are not necessarily high in mechanical aptitude. Just as in other subjects, individuals who enter courses in manual arts should have aptitude for them.

Tests of mechanical ability are used both to measure school achievement and to indicate the presence of mechanical aptitude. More than is the case with tests in other fields, prediction is an important function of the test. These instruments of prediction foretell the probable success of a student not only in the manual arts but also in the occupation which he is most likely to enter.

The school's function is to acquaint the students with the breadth and significance of this area which fills such a large place in our civilization. This is possible through trips, reading, and descriptions on the one hand and through participation in some actual occupation on the other.

Well-planned courses in industrial and practical arts strive to fulfill this need.

Courses in manual arts in the elementary school are apt to be rather general in nature, with less emphasis on precision in constructed objects and more upon a general understanding of the part that practical and industrial arts play in our civilization. The materials of the course frequently grow out of the problems being faced daily by the members of that community. Their main purpose is exploratory in that the child explores his interests, aptitudes, and general fitness for occupations which require the coordination of mind and hand. Such a one who makes a table or a lampstand appreciates more keenly the work required to construct an acceptable commercial object and consequently is more apt to acquire a new respect for labor and the laboring man. These courses in the manual arts, then, are characterized by a considerable variety because they vary with the environment in which the school is placed.

In the junior high school there is also a wide differentiation among courses. Boys' aptitudes are provided for in such courses as manual training, plumbing, electricity, woodworking, metalworking, cabinet-making, etc., while those of girls are met in domestic science, household arts, prenursing, bookcraft, or home decoration. These courses require more exactness in the objects constructed and more workmanlike form in the processes used. Because there is such a rich variety of courses, very few standardized tests have been widely used. Tests of information are rather easy to construct, but standardized tests or scales for use in judging more exactly the objects made in these courses are few indeed.

OBJECTIVES IN THE TEACHING OF MANUAL ARTS

As we have often said in the course of this text, the objectives must be clearly defined before a satisfactory test can be constructed. The general outcomes of courses in manual arts can be briefly stated:

1. To furnish the student with wide experiences in industrial and practical arts. In this manner he can discover something of his own interest and aptitude for that sort of work. Thus a child who is contemplating leaving school may reconsider when he engages in the actual construction of some object which he wants personally. Such an interest may become permanent and give direction to his whole afterschool life.
2. To develop an appreciation of the world of manual work: (a) to furnish experiences of common value, shared by all who take the work, so that sympathetic attitudes may be developed toward other workers; (b) to develop also some actual skills in mending and improving mechanical gadgets around the home, and (c) to furnish an insight into the quality of those articles which need to be purchased.

3. Finally, to (a) furnish an opportunity to develop special aptitudes, (b) stimulate a need for further courses by pointing out the part that science and mathematics play in successful industrial work, and (c) offer special work such as printing to those who are soon leaving school to go to work.

Many of these outcomes of instruction in the manual arts have not as yet been measured. Easiest of all to measure is the amount of information possessed. Interest in mechanical activities is well reflected in the scores of our interest inventories as developed in Chap. 16. Measures of aptitude also will be described in the course of the present chapter.

TESTS

Nearly all the tests in fine arts contribute something to the measurement of industrial arts. Especially is this true of the McAdory Art Test and the Lewerenz Tests in the Fundamental Abilities of Visual Arts. Among the tests of woodworking and mechanical drawing, only the Nash Van Duzee Industrial Arts Tests will be described. The Nash Van Duzee Industrial Arts Tests¹ are divided into two tests:

Test I. Woodwork

Scale A. Technical and related information

Scale B. Performance

Test II. Mechanical drawing

Part I. Information

Part II. Performance

Scale A of Test I is composed of true-false items. Multiple-choice items with three choices test the processes and methods used in woodwork, the care and use of basic hand and machine tools, etc. Test I also uses diagrams to test knowledge and understanding of common joints used in woodwork, and incomplete drawings of a simple wood block to test the pupils' understanding of a drawing as "to placement of views, methods of representing shapes in shop drawings," etc.²

Scale B consists of an actual piece of wood and the proper tools with which certain processes are to be performed according to a working drawing. The subject has for example to "plane, square, and true" (1) a face, (2) an edge, (3) an end. He must among other things plane the chamfer straight and true and chisel the mortise smooth. A booklet is furnished which aids the tester in scoring the details of the performance.

Test II, Mechanical Drawing, is made up of processes which an investigation in many schools proved to be generally used. As in the preceding test, Part I consists of completion and multiple-choice tests

¹ Bruce Publishing Company, Milwaukee. Item by permission.

² *Manual*, Test I, Woodwork, Scale A.

of information relative to mechanical drawing. It also has a test of interpretation of the conventions of drawings and machine drawings. Part II contains tests of dimensioning, geometrical constructions, making a working drawing, lettering, and orthographic drawing (Fig. 22).

The reliability of the tests varies from .61 to .94 with a median about .87 for the test as a whole. Its norms are unique indeed, for instead of the usual median or percentile for each grade, the norms of median and best score are given for the number of minutes the course has been

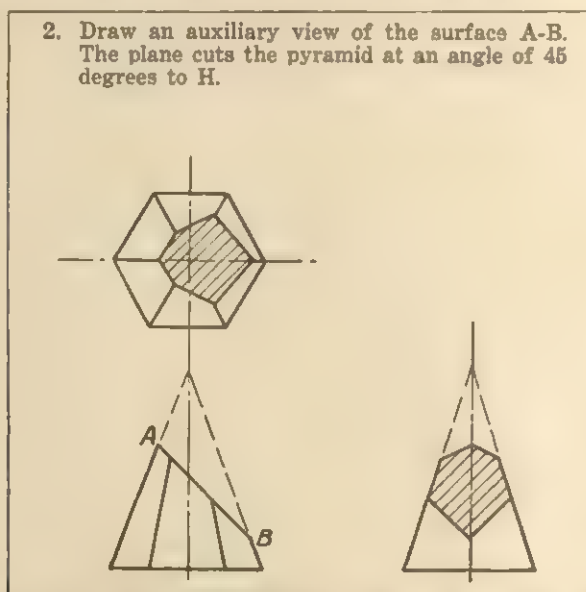


FIG. 22. Section D, Orthographic Drawing, Nash-Van Duzee Industrial Arts Tests.

studied. The best score refers to the median performance of the best school studied. Tables are presented also for changing scores into school marks. The validity of the test rests on the care with which courses of study were investigated in constructing the items of the test. Altogether this is a very satisfactory instrument for measuring the outcomes of courses in woodworking and mechanical drawing.

HOME MECHANICS

The construction and standardization of a test of mechanical achievement are well illustrated in the Newkirk-Stoddard Home Mechanics Test. The procedure is sound because the objectives and materials of the course grew out of an investigation of the uses of mechanical devices within the home.

A study was carried on to determine a list of the practical jobs of a mechanical nature ordinarily done around the home. First, 382 home jobs of a mechanical nature were reduced to 130 which were practical and were adapted to shop instruction. Then description of these jobs was sent through a questionnaire to "100 mature people who have homes in the middle west," who were asked to check the jobs which they had occasion to perform. The investigation also sent a questionnaire to a number of schools to discover what was being taught in their courses in home mechanics; 75 schools replied. Altogether 72 home-mechanics jobs were selected (1) because they were widely used in home and mechanics courses, and (2) because they stood high in social utility. The test, then, went through an experimental and a final edition. Two forms, with 36 jobs in each, were constructed. A composite table of percentages of accomplishment shows the percentages of achievement in grades 7 to 9 in each of 10 schools. The reliability of the test is not emphasized, but it correlates .44 with the Otis Intelligence test, .26 with the Stenquist Assembly Test, and .64 with teachers' marks in a course in eighth-grade home mechanics.¹ The two following examples from the test illustrate the type of activities which compose the test and the manner in which measurement was made. The directions state that all procedures are to be rearranged in the right order according to their numbers.

8 To make a joint with tinner's rivets

1. Head the rivets
2. Get the seam in place
3. Set the rivets
4. Spot the rivets

() () () ()

21 To assemble a radio set

1. Wire the set according to circuit diagram
2. Secure the necessary parts and supplies
3. Decide on a circuit
4. Mount instruments on panel and baseboard
5. Drill the panel and fasten to baseboard
6. Lay out panel and baseboard

() () () () () ()

This test is introduced here more as a sample of procedure than as a useful standardized test. In the first place, the samples of mechanical work in the Middle West might not be the same as those in the East, South, or Far West. Nor could the norms be applied in other sections of the country without modification. On the other hand, the procedure in test construction which discovers what is actually being done in the home and then checks this outcome with the school procedure is sound.

¹ Newkirk, Louis V., *Validating and Testing Home Mechanics Content*. Studies in Education, Vol. 6, No. 4. University of Iowa, 1930-1932. Items by permission.

HOME ECONOMICS

The objectives of instruction in home economics may be divided into the immediate and the more remote. Immediate objectives are:

1. To develop skill in the selection, preparation, and serving of foods. This involves the acquisition of (a) the knowledge and understanding of the facts and principles of nutrition, as well as (b) the application of these facts and principles to the actual preparation of food for the table.

2. To develop efficiency in exercising good judgment in the selection and making of clothing. This efficiency depends upon the acquisition of information about the characteristics of different kinds of cloth, about the use of patterns in cutting out garments, and the aesthetic effect upon the person of different kinds and colors of cloth arranged in a variety of ways in garments, etc.

3. To understand the characteristics which make for an efficiently run household. In this division there are problems of the proper management of time and money, of good social relations within the household as well as of house planning, of house furnishing, and of house care.

4. To understand and to apply to the care of the home the best principles of aesthetics, hygiene, and sanitation.

The More Remote Objectives Aim to develop within each individual those attitudes which will result in consideration for the comfort and convenience of others as well as in a willingness to serve for the common good of the whole family.

Measurement in Home Economics

Most easily measured are the facts which an individual possesses about foods, clothing, and management of the household. Most difficult to measure are the eating habits which an individual practices and the success he has, for example, in making pies. Objective tests are customarily administered to test facts of information; check lists and rating scales, for performance.

The Engle-Stenquist Home Economics Test has served as a useful instrument in this field since 1931. It included suitable items concerned with foods and cookery, with clothing and textiles, and with household management which were intended for grades 5 to 10. But it became old and out of date and is now out of print.

In addition, a series of tests for several branches of home economics have been prepared at Purdue University.¹ The accompanying table lists the titles of four tests, all suitable for grades 7 and 8.

¹ State High School Testing Service, Purdue University, Lafayette, Ind. Items by permission.

| Test | Time, minutes |
|--|---------------|
| 1. Assisting with Clothing Problems..... | 28 |
| 2. Helping with the Housekeeping..... | 28 |
| 3. Helping with Food in the Home..... | 28 |
| 4. Assisting with Care and Play of Children..... | 28 |

While these tests are based on the course of study of the state of Indiana, they deal with common principles. These tests have no published norms or reliabilities but deserve mention because they cover each unit thoroughly. They furnish highly suggestive techniques for tests in home economics for grades 7 and 8.

Tests of Home Economics: High School

Measurement in the field of home economics at the high school level is divided into two parts:

1. Tests of information in the areas of food, clothing, and home making
2. Rating scales (a) of habits and procedures used in preparing food, and (b) of the foods themselves.

In the tests of information and understanding we must turn again to the tests prepared by a group of workers for the state of Indiana. The accompanying table lists the tests.

| Test | Time, minutes |
|--|---------------|
| 1. Clothing I..... | 55 |
| 2. Clothing II..... | 55 |
| 3. Foods I, Food Selection and Preparation..... | 55 |
| 4. Foods II, Planning for Family Food Needs..... | 55 |
| 5. Child Development..... | 55 |
| 6. Home Care of the Sick..... | 55 |
| 7. Housing the Family..... | 55 |

These tests are carefully constructed, cover the areas well, and test for both information and understanding. They are not, however, standardized tests because they have no norms or computed reliabilities and are still in the mimeographed stage. A more detailed description of one of these tests will give an idea of the soundness of the above statements.

The test titled Foods I. Food Selection and Preparation, contains 175 items to be answered by + or 0 (true or false), multiple choice, and matching. Sometimes the items are couched in the form of a problem or situation, as, for example, the presentation of a menu which is to be evaluated by checking whether it contains a variety of color, is a fuel-saving meal, contains little starch, etc. Here is one illustration of a matching problem:

Place in the blanks at the right of Column II the letter of the food group in Column I that best identifies the item in Column II or the function in Column II. The first question is done correctly to show you how to proceed. Some items may be used twice and some not at all.

| <i>Food Groups—Column I</i> | <i>Items—Column II</i> | |
|-----------------------------|--|---------------------|
| (a) protein foods | 85a. Meat, poultry, and fish | <u> </u> a 85a |
| | 85. Codliver oil | <u> </u> 85 |
| (b) Vitamin A | 86. Calcium, iodine | <u> </u> 86 |
| | 87. Sugars and starches | <u> </u> 87 |
| (c) minerals | 88. Butter, shortening, bacon fryings | <u> </u> 88 |
| (d) Vitamin C | 89. Found in green and yellow vegetables | <u> </u> 89 |
| (e) carbohydrates | 90. Bread, potatoes, cereals | <u> </u> 90 |
| (f) vitamins | 91. Found in sunshine | <u> </u> 91 |
| (g) fats | <i>Functions—Column II</i> | |
| | 92. Prevents "night blindness" | <u> </u> 92 |
| (h) Vitamin D | 93. Repairs body tissues | <u> </u> 93 |
| | 94. Prevents rickets in children | <u> </u> 94 |
| (i) sugar | 95. Provides energy quickest in the body | <u> </u> 95 |
| (j) starchy foods | 96. Necessary for healthy teeth and gums | <u> </u> 96 |
| (k) Vitamin B ₁ | 97. Necessary for growth | <u> </u> 97 |

Many other problems are included, such as what Mary needs to do in preparation for the family breakfast, what foods should be eaten by high school students, why Edith's cake fell in the center, and what correct practices Barbara observed at a dinner party. In general, the list of answers is furnished, the student checks the correct one.

In the ratings of habits and of foods, the Minnesota Check List for Food Preparation and Serving by Clara M. Brown¹ consists of 13 rating

| | 1 | 2 | 3 | 4 | 5 | Score |
|----------------------|--|---|---|---|---|-------|
| 1. Grooming | Untidy; hands or nails dirty; dress soiled or inappropriate, no apron; hair in disorder and unconfined | Reasonably well groomed; dress suitable; apron soiled or wrinkled; hair neat but not held in place | | Immaculately clean; dress and apron fresh, unwrinkled and appropriate; hair held in place by band or covering | | (1) |
| 10. Setting of table | Wrong dishes, silver, or table cover used or arranged incorrectly; table looks crowded | Dishes, silver, and table cover suitable and arranged correctly; centerpiece lacking or inappropriate | | Dishes, silver, and table cover suitable and correctly arranged; decorations attractive | | (10) |

¹ University of Minnesota Press, Minneapolis. Items by permission.

scales. Each of the 13 traits rated is accurately described at three levels of achievement. Two of the scales are shown in the preceding table. One simply checks each of the 13 scales at the point which describes the subject and adds up the points. The value of using a check list is clearly stated in the manual:¹

Value of Using A Check List

The following statements of the results of using a check list are based upon the findings of experimental studies.

1. *Learning proceeds more rapidly when goals are clearly defined* than when the learner has only a vague idea regarding them. Use of the check list enables students to see clearly what desirable standards are.

2. *Pencil-and-paper tests, no matter how high a degree of reliability they possess, are not valid measures of a person's ability to do certain tasks.* The correlation between knowledge as recorded in pencil-and-paper objective tests and the abilities listed in the check list appears to be considerably below .50. Since this is true, it is essential that standards of appearance, personal habits, and work abilities be evaluated if these are regarded as important goals.

3. *Providing descriptions of low and average achievement* as well as of the high level increases accuracy of rating and enables students to understand wherein they fail to reach the standard.

4. *Objective self-evaluation tends to accelerate the rate of learning,* and the use of such devices as the Minnesota Check List permits individuals to judge their own achievements and limitations.

There are no norms, or published reliability, or any correlations of the results with other criteria.

The final rating instrument here described is the Minnesota Food Score Cards, revised edition² which was constructed under the direction of Clara M. Brown. These cards contain rating scales for judging the quality of 57 foods. The precise wording of the rating scales increases the objectivity of scoring. The food score cards are prepared for such foods as bacon, coffee, eggs (five kinds), fruit cup, piecrust, popovers, candy (four kinds), soufflé, tea, and waffles. These cards are especially constructed to rate the success of students in actually preparing food in the laboratory. One example is shown in the table on page 316.

¹ University of Minnesota Press, Minneapolis. By permission.

² Cooperative Test Division, Educational Testing Service, Princeton, N.J. Item by permission.

ICE CREAM

| | 1 | 2 | 3 | Score |
|----------------|--|-------------------------------------|---|-------|
| Color..... | 1. Muddy or pale | Clear and uniform | | 1. |
| Consistency... | 2. Too hard or runny | Just firm enough to hold shape | | 2. |
| Texture..... | 3. Coarse, granular, or fluffy | Smooth, velvety, compact | | 3. |
| Flavor..... | 4. Flat, insipid, or too highly flavored | Delicate yet definite; well-blended | | 4. |

MECHANICAL APTITUDE AND ABILITY

Thus far we have considered achievement tests in the fields of fine arts, mechanical arts, and home economics. The rest of the chapter will be devoted to a consideration of the measurement of mechanical aptitude or ability.

USES OF TESTS OF MECHANICAL ABILITY

Mechanical-ability tests have two outstanding spheres of usefulness. The first of these deals with the ability of the student to profit by courses involving mechanical ability. Paterson, for example, showed that the Minnesota Mechanical Assembly Test correlated more highly (.53) with the final marks in such a course than a test given in the first half of the course (.42). Thus a test given in 1 hour's time predicted final standings in the course better than 6 weeks of experience. In the second place, tests of mechanical ability are directly correlated with subsequent success in a variety of occupations which utilize mechanical processes and information and hence are useful for vocational-guidance purposes. For example, the two-hand test of mechanical ability, which consists of the control of the direction of a pointer by two screws which work at right angles, correlates .57 with machine operating, .59 with toolmaking, and .62 with turning (lathe work).¹

PROCEDURES USED IN TESTING

There are three procedures which may be used to test mechanical ability: (1) analyze the mechanical processes into simplest elements and test them, (2) construct tests of information which sample the types of mechanical information accumulated up to that time, and (3) disarrange or strip a set of mechanical gadgets and have the student assemble them.

¹ Bingham, Walter Van Dyke, *Aptitudes and Aptitude Testing*, p. 135. New York: Harper & Brothers, 1937.

Analysis of Processes into Elements

Just as the Seashore test of musical ability may divide this ability into pitch, intensity, time, rhythm, etc., in like manner mechanical ability may be analyzed into (1) reaction time, (2) agility and strength, (3) manual dexterity, (4) steadiness, (5) manual rhythm, etc. To measure these abilities efficient measuring instruments have been constructed. Reaction time has been measured by a chronoscope in thousandths of a second.

Reaction time is the elapsed time between the giving of a signal and the performance of some defined act. Under the simplest conditions an individual sits with one hand on a telegraph key which he pushes down whenever a light is flashed. The signal may be a flash of light, a sound, a taste, a touch, a smell, pain, etc. The reaction time depends on such things as the set of the individual, the intensity of the stimulus, and of course the type of individual. When such measures are made on the same individual we find large differences in the reaction time which depend upon the modality employed. For example, the reaction time for a touch on the hand averages about 0.120 second, while it takes 1.082 seconds to respond to a bitter taste.

Several simple measures of simple abilities are now considered. Agility of young children has been measured by their capacity in jumping, catching balls, and climbing ladders. Manual dexterity has been measured by simple tapping in which an individual strikes a brass board as rapidly as possible with a stylus which is in circuit with a counter. The counter registers each tap. Steadiness is measured by a subject's moving a brass stylus between two converging brass plates until he touches one of them or else by putting a stylus into holes graduated in size without touching the sides of the hole. Rhythm has been measured by having an individual listen to a sequence of four notes which is repeated for several times. The test comes in keeping time with the sequence by pressing a telegraph key.

There is no question about the accuracy of these measurements. They do well what they purport to do, but they do not correlate with or predict the ordinary mechanical performances with which the school is concerned. These latter activities are much more complex and include these simpler functions in a great variety of combinations.

Tests of Information about Mechanical Ability

In these tests many types of information about mechanical devices and processes are sampled. The assumption is that those individuals who have good mechanical ability will be continually examining the machines which are around them, will read accounts of new machines

in such magazines as *Popular Mechanics*, and thus will accumulate mechanical information. On the other hand, those possessing little mechanical ability will not examine machines nor will they care to read about them and so they will not accumulate information on machines and their processes. Unfortunately for the use of this criterion, it is substantially correlated with intelligence and hence does not furnish a unique measure of mechanical ability. Here are a few examples from the Detroit Mechanical Aptitudes Examination for Girls:¹

- | | | | | |
|-----------------------------------|----------------|-----------------|-----------------------|-----------|
| 14. Solder will stick best to | 1 glass | 2 lead | 3 leather | 4 wood |
| 20. Glass is usually cut with a | 1 chisel | 2 files | 3 scissors | 4 wheel. |
| 23. A spark plug is in the | 1 commutator | 2 cylinder head | 3 manifold | |
| | 4 piston. | | | |
| 27. A carburetor | 1 explodes gas | 2 measures gas | 3 mixes air with gas. | |
| 35. An electric doorbell requires | 1 current | 2 fuse | 3 plug | 4 switch. |

In addition, practically all paper-and-pencil tests of mechanical ability have one or more sections which are dependent upon mechanical information for their correct answers.

Mechanical Assembly and Performance Tests

Mechanical assembly tests, as their name implies, consist of putting together in the correct manner parts of disassembled mechanical gadgets. Stenquist's original mechanical assembly test was made up of such objects as a bicycle bell, a chain with split links, a small door lock, and a mousetrap. The disassembled parts were to be reassembled by the aid of a screwdriver. An assembly test was also constructed by Toops which contained items lying more nearly in the usual environment of girls. Such problems as the stringing of beads, cross-stitching, tape sewing, card wrapping, and making a trunk tag were used. All these assembly tests demanded a great variety of psychological processes including perception, steadiness, and manipulation. Since they were more like real-life situations in mechanical performance they tested well some aspects of mechanical aptitude. Many of the tests later to be described contain aspects of these three types of measurement.

Assembly tests of mechanical ability demand some sort of performance for success but differ greatly in the type of material utilized. Only a few tests will be mentioned here. Among performance tests, we shall discuss (1) the Minnesota Mechanical Assembly Test, and (2) the MacQuarrie Tests for Mechanical Ability. Among paper-and-pencil tests, we shall discuss (1) the Revised Minnesota Paper Form Board, (2) the Mellenbruch Mechanical Aptitude Test for Men and Women,

¹ Items by permission of Public School Publishing Company, Bloomington, Ill.

(3) Aptitude Tests for Occupations, and (4) the Differential Aptitude Tests.

Performance Tests

Of this group, the Minnesota Mechanical Assembly Test is of the first importance. The builders of this test first made a thorough canvass of the available tests.¹ Among the many tests investigated, the Stenquist Mechanical Assembly Test proved satisfactory save in one important particular, it had a low reliability. This rather short test was lengthened. New items were tried out and the successful ones embodied in the test until there were three boxes -A, B, and C—each of which contained 11 gadgets to be reassembled (Fig. 23). You will note that this test contains such gadgets as a large paper clip, an ordinary lock, a safety razor, a pair of pliers, scissors, a bicycle bell, a die holder, an expansion nut, and many other mechanical objects to be put together.

As in all other tests, the most difficult problem of all was the establishment of the test's validity. In so many cases the criterion against which we measure the validity of a test is no more sound than the test itself. In the present instance a criterion was desired which had in it the essence of mechanical ability. The criterion finally selected was the quality of the mechanical work actually produced in a junior high school class of mechanical arts. Every effort was made to measure accurately products of the class's workmanship. In the first place *direct observation* and inspection were made as to whether, for example, letters were transposed in printing, whether the working lines showed in manual drawing, whether there were loose wires in electrical wiring or parts chipped in woodworking. In the second place, *actual measuring devices* were applied whenever possible. Rulers were used to measure distances, calipers to measure dimensions in mechanical drawing, stencils to measure rounded corners, steel square to locate rivets, and a graduated small wedge to measure the flatness of boards. In the third place, *scales* were constructed with graduated samples of increasing fineness of quality. There was thus one scale for rating the soldering of biscuit cutters, another for judging the splices of wire in electricity, and another for judging lettering.

The results of these three criteria were combined in the optimal manner to obtain a quality criterion which was reliable and dependable and against which all tests could be measured. It was seen that three tests stood out above the others in their correlations with this criterion. The Minnesota Mechanical Assembly Test correlated .55 with this criterion, the Minnesota Spatial Relations Test, .53 and the Minnesota Paper

¹ Paterson, Donald G., *et al.*, *Minnesota Mechanical Ability Tests*. Minneapolis: University of Minnesota Press, 1930.

Form Board, .52. When the results of an information test of mechanical processes were added to this criterion of quality only the paper form board was increased substantially (.52 to .65). It was shown that the criterion of information was correlated with intelligence and hence added little of significance to the measurement of mechanical ability.



FIG. 23. Materials from Minnesota Mechanical Assembly Test, short form, Boxes I and II. (By permission of the Marietta Apparatus Company, Marietta, Ohio, and Professor Donald G. Paterson.)

The MacQuarrie Tests for Mechanical Ability are not assembly but performance tests. According to the manual more than five million persons have had their mechanical aptitudes assayed by this test.

There are seven tests in the battery. Three of them, tracing, tapping, and dotting, have a large manual-dexterity element. Tracing consists in drawing lines through small openings placed in a series of vertical lines about $\frac{1}{8}$ inch apart. Tapping consists simply of putting three pencil dots

as fast as possible in a series of circles, all of equal size and the same distance apart. In the dotting test the subject places one dot in each small circle in a connected line of circles occurring at irregular intervals. The second group of tests consisting of copying, location, blocks, and pursuit - are more closely related to intelligence. The copying test consists of tracing out on dots arranged in rectangular order a simple figure. The point of beginning is indicated with a circle around the proper dot. The test of location consists of recognizing on a smaller area the position of letters placed in a much larger area. In the blocks test a set of blocks drawn all the same size are piled up in a variety of ways. The problem is to count by direct visual inspection and visual projection the number of blocks which touch a marked block. In the pursuit test,

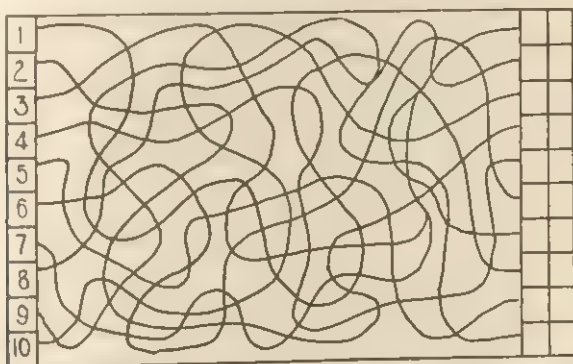


FIG. 24. Pursuit. (By permission of T. W. MacQuarrie, Professor of Education, University of Southern California.)

the eye must follow a single wandering line through a maze of other lines to its correct destination (Fig. 24).

Norms have been computed for ages 10 to 16 and for adults. These latter norms are based upon 1,000 males and 1,000 females, aged 16 and up. The reliability and validity of the test are discussed in the succeeding paragraphs.

What characteristics have led to such wide application? The characteristics of brevity, ease of administration, separateness of the subtests, and success in prediction have recommended it. The test can be administered in 30 minutes. It is easy to give and score and its norms are satisfactory. Its reliability was computed both for the subtests taken separately and for the battery as a whole. The reported reliabilities are shown in the table at the top of page 322.

Thus not only is it possible to correlate scores on the test as a whole with success in any occupation, but any single test's score, or any combination of scores weighted in any manner, may be likewise correlated.

| Test | Reliability |
|------------------|-------------|
| 1. Tracing..... | .80 |
| 2. Tapping..... | .75 |
| 3. Dotting..... | .74 |
| 4. Copying..... | .86 |
| 5. Location..... | .72 |
| 6. Blocks..... | .80 |
| 7. Pursuit..... | .76 |
| Total Score..... | .90 |

One study (Harrell and Faubion, 1940),¹ for example, concluded that the tracing subtest predicted more accurately the elements of metalwork than the test as a whole. It is thus possible to use optimum weights for each prediction.

At Hunter College, it was demonstrated that a combination of pursuit, tracing, and dotting predicted success in typing. Lawshe pointed out that a multiple *R* with optimum weighting correlated .46 in selecting radio-assembly operators, while the total test's correlation was .42.² Its correlations with success records in occupations have been keys both to its use and to its validity. It has been correlated with such mechanical occupations as aviation mechanics, aircraft inspectors, machinists, tool-maker apprentices, gun wrapping, and mechanical drawing. While the correlations with these criterion scores have rarely been as high as .50, they have shown their worth in combination with other predictive factors.

The MacQuarrie test has also proved its value in predicting success in high school in mechanical drawing as well as in projects of construction. In one high school, the test showed a significant difference between students judged to be most promising and most unpromising.³ Moreover, in another study, where pupils aged 12 to 15 developed a project in electrical construction, the correlation between test scores and accomplishment in the project was .79, in time to complete the project, .72.⁴

MacQuarrie's definition of mechanical ability throws some light on the nature of his test. "Mechanical ability," he writes, "is broadly defined as a pattern of specific aptitudes such as eye-hand coordination, speed of finger movement, and ability to visualize space."⁵ The test

¹ Harrell, Willard, and Richard Faubion, "Selection Tests for Aviation Mechanics," *Journal of Consulting Psychology* (1940) 4:104-105.

² Buros, Oscar K. (ed.), *The Third Mental Measurements Yearbook, op. cit.*, Item 661, p. 690.

³ Stoy, E. G., "Additional Tests for Mechanical Drawing Aptitude," *Personnel Journal* (1928) 6:361-366.

⁴ Horning, S. D., and Ruth S. Leonard, "Testing Mechanical Ability by the MacQuarrie Test," *Industrial Arts Magazine* (1926) 15:348-350.

⁵ *Manual*, p. 1.

itself consists of a set of tests to measure specific aptitudes. It undoubtedly emphasizes manipulative skills which involve the dexterity of finger and hand, acuity of vision, the control of muscles, and the perception of space. There is little in the test concerned with the understanding of the fundamental principles of mechanics or with familiarity with the common tools. Like other tests which predict, this test is plagued with low correlations. How can real prediction be much better than chance when the predictive instrument correlates .45 with the criterion of success? Remember, the efficiency of a correlation coefficient of .45 is just 11 per cent better than chance. Unless many other factors are used in the prediction, a counselor will go wrong much more often than he will go right when he uses such an instrument.

Paper-and-pencil Tests

In the paper-and-pencil tests, indications of the presence of mechanical ability are secured through tests of information, by matching of pictures of objects which in some way belong together, and by figuring out what the result would be from a pictured situation. While these tests differ widely in their content, they are all alike in requiring no physical manipulation of machines or any particular performance beyond that of putting down the answer. Four tests are reviewed here: (1) the Revised Minnesota Paper Form Board, (2) the Mellenbruch Mechanical Aptitude Test for Men and Women, (3) Aptitude Tests for Occupations, and (4) the Differential Aptitude Tests.

The Revised Minnesota Paper Form Board Test was an outgrowth of the Army Beta and a more complicated form board of O'Rourke. The present edition requires the subject to recognize out of five single drawings that figure which represents the two figures which are separated. Three illustrations (Fig. 25) will make clear the nature of the test.¹ These illustrations show that the problem here is to discriminate patterns in two dimensions. Studies show that the test correlates .25 to .30 with grades in descriptive geometry; .40 to .45 with some of the semi-skilled occupations and .57 with test scores and success of inspector packers.² Some investigators found the test of less value in these various occupations because it had low correlations with intelligence scores and low correlation with mechanical-aptitude tests.

The manual claims, and with some justification, the following:³ "The evidence thus far accumulated appears to indicate that high scores on this test are predictive of (1) ability to learn mechanical draw-

¹ Items by permission from the Psychological Corporation, New York.

² See Stuit, Dewey B., *The Third Mental Measurements Yearbook* (Oscar K. Buros, ed.), Item 677.

³ *Manual*, p. 2.

ing and descriptive geometry; (2) success in mechanical occupations; and (3) success in engineering courses." They base their contention about the geometry on a correlation of .25 to .30 (certainly not too solid a base) and their prediction about engineering on the fact that engineering students scored higher on the test than did others. Success in mechanical prediction rests on a study by Crawford (1941)¹ which indicated that this test was superior to others in predicting mechanical ability. The reliability for one form is reported as .85 and for both together as .92. Norms based on a heterogeneous population of 5,000 subjects are available. The revised edition is machine-scored, but the norms for this edition are based on only 548 white enlisted men. Because

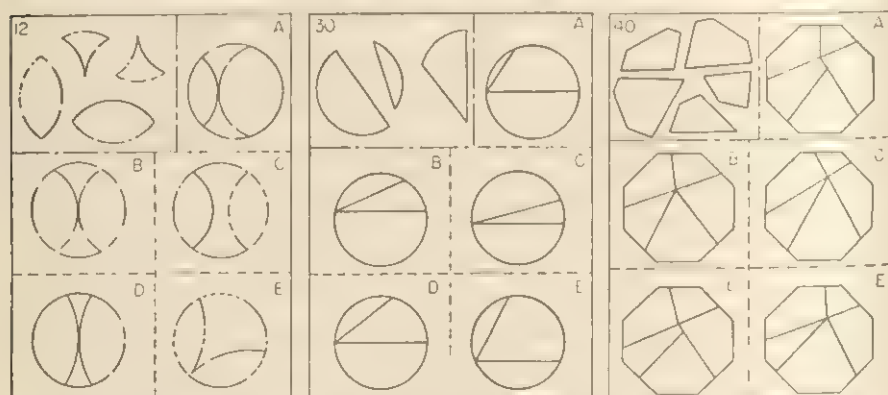


FIG. 25. Revised Minnesota Paper Form Board Test, Items 12, 30, and 40

of the ease of administration and the small amount of time needed to give the test (20 minutes), the Revised Minnesota Paper Form is one of the most widely used measures for testing the components of engineering and mechanical aptitude.

A second test of the pencil-and-paper variety is the Mellenbruch Mechanical Aptitude Test for Men and Women. This test is applicable from grade 7 through adulthood. The test consists of seven sets of 12 pictures. On one side of the page the mechanical objects are numbered. On the other side they are lettered. The problem is to match the letters with the proper numbers. Figure 26 is a sample page. Considerable care was exercised in constructing the tests. An original list of 425 paired photographs were tried out against ratings for workers in machine shop, sheet metal, woodworking, blueprint reading, and mechanical drawing. Items were selected which correlated well with these criteria and which did *not* show a sharp distinction between boys and girls or between men

¹ Crawford, John Edmund, *Measurement of Some Factors upon Which Is Based Achievement in Elementary Machine Detail Drafting*, unpublished doctor's thesis, University of Pittsburgh, 1941.

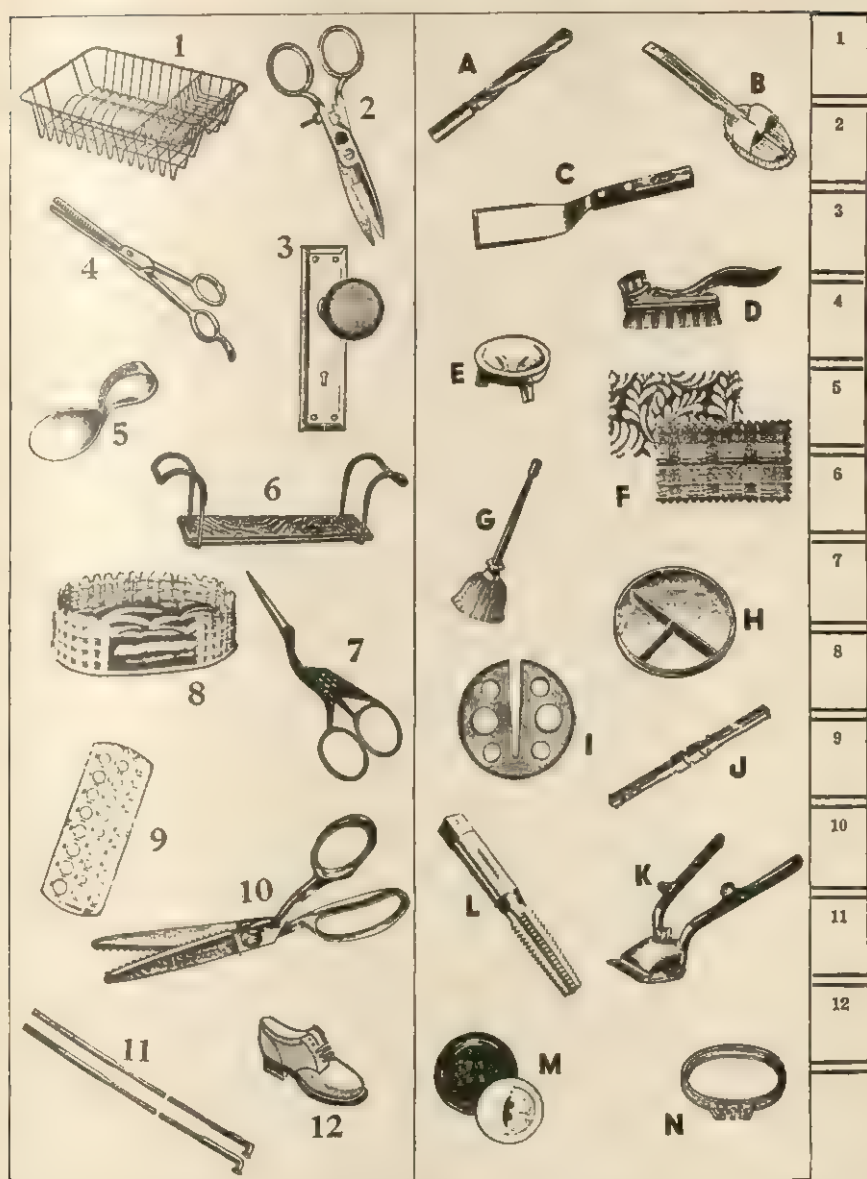


FIG. 26. Page 3, Mellenbruch Mechanical Aptitude Test for Men and Women. Match letters with numbers. (By permission of Paul L. Mellenbruch.)

and women. The tests which showed differences between the sexes were discarded so that the test as a whole shows only a 6-point difference in favor of boys for comparable ages. This characteristic, however, may be a weakness because so many tests have shown such a decided difference between boys and girls in this capacity that the difference may be a real fact. At any rate, this test can be successfully used for both boys and girls, both men and women.

The Mellenbruch Mechanical Aptitude Test is well constructed, is easily administered to groups of students, and has satisfactory reliability and fair validation. The correlation of Form A with Form B is .87. Some light is thrown upon its validity by correlations of various kinds which are reported in the manual. The coefficient between the test's scores and teachers' ranks in a course of engineering drawing was .57 (57 women) and with the degree of participation in mechanical activities of 430 unselected men and women was .60. It also correlates well with other measures of mechanical ability. A test of mechanical ability must correlate low with intelligence or else it is just another test of intelligence. In this case the correlation coefficients ranged from .17 to .33 which are low enough to be satisfactory. Satisfactory norms are provided for Grades 7 to 12, college freshmen, and a wide range of mechanical occupations.

Two uses are clearly indicated for this test: (1) to help decide whether a student would profit from courses in manual arts, and (2) to indicate an individual's aptitude for those occupations which require a considerable amount of mechanical ability. The manual recommends as follows:

1. That an individual who receives fewer than 30 points on the test be not employed for mechanical work.

2. That an individual who receives 30 to 40 points be employed for simple routine manual tasks.

3. That an individual who receives 40 to 55 points be employed to perform complex but routine tasks.

4. That an individual who receives above 55 points be employed to perform tasks demanding mechanical ingenuity.

In the third place, the test of mechanical aptitude, Form A, is the second of six Aptitude Tests for Occupations.¹ It consists altogether of pictures and drawings and contains the following items:

| | Number |
|--|--------|
| 1. Objects or tools and their use..... | 19 |
| 2. Patterns which represent objects or are to be used in their construction. | 19 |
| 3. Patterns that fit designs..... | 8 |
| 4. Motor driven shafts and pulleys..... | 7 |
| 5. Names of joints..... | 7 |

¹ California Test Bureau, Los Angeles, Calif. By permission.

Twenty minutes are allowed to take the test and the answers are recorded on a separate sheet. Examples are shown in Fig. 27. This test has been partially *validated* by correlating it with other tests of mechanical aptitudes and with courses in machine shop (.40) and mechanical draw-

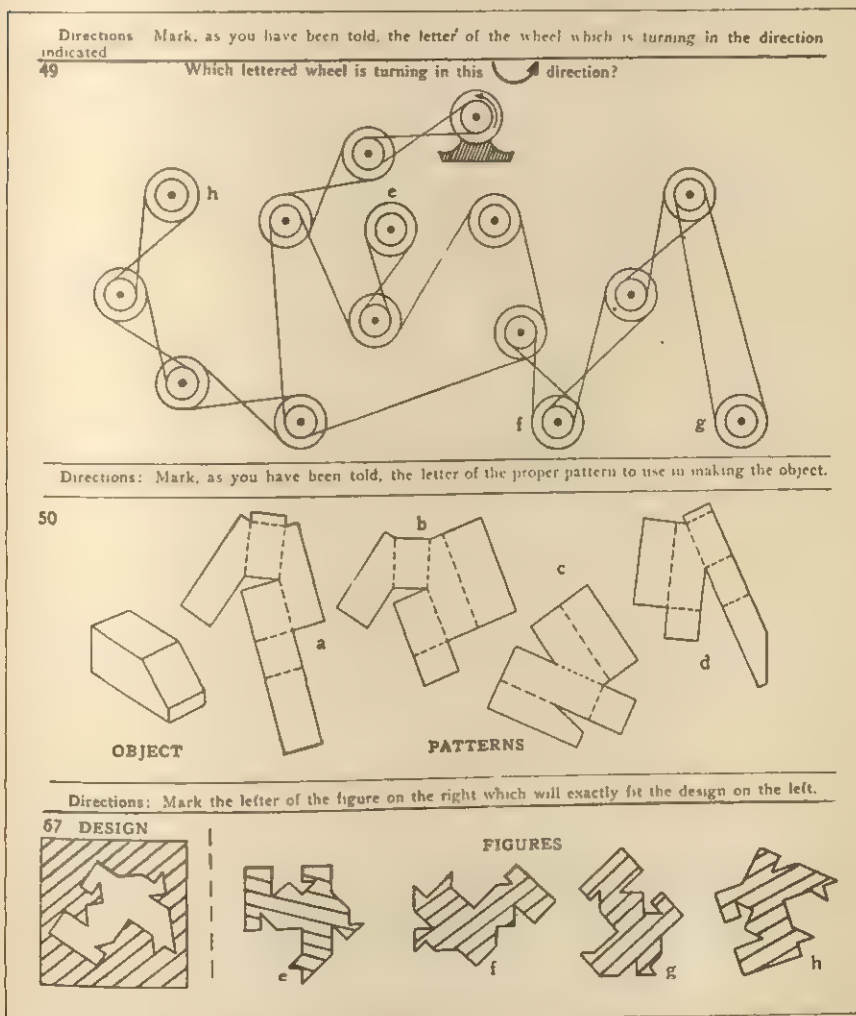


FIG. 27. Aptitude Tests for Occupations, mechanical aptitude. (Roeder and Graham.)

ing (.35). With tests of mechanical knowledge, mechanical comprehension, spatial relations, and mechanical ability the correlations range from .41 to .64 (manual). The *reliabilities* for men and boys range from .76 at age 13 to .83 at age 9. For girls and women the coefficients are somewhat lower. As a whole the test is probably too short for high

reliabilities. There is a chance that increase in the time allowed would improve the test.

In the fourth place, the Differential Aptitude Tests¹ have much to recommend them. They are suitable for grades 8 to 12. These eight tests, except for the clerical test, are power tests, *i.e.*, the items increase in difficulty. They are all standardized on the same population thus making their percentiles comparable and enabling the teacher to make a profile from the scores. Their instructions are clear and their scoring done either by hand or by the IBM machine. Percentile norms for both sexes based on over 20,000 well-selected cases are available for grades 8 through 12. These aptitude tests are as follows:

1. Verbal Reasoning is composed of 50 analogies with their extremes omitted. There are four choices for the first part and four choices for the second. It takes 30 minutes to give and its reliability is .90.

2. Numerical Ability includes 40 examples to be worked. These include the subtraction, addition, multiplication, and division of simple numbers, common and decimal fractions, and mixed numbers. It includes items involving square root, cube root, and proportion. It takes 30 minutes for administering and has an average reliability of .90.

3. Abstract Reasoning consists of 50 sets of drawings. One must pick out the pattern developed among four drawings from among five choices in the answer. Time for administering is 25 minutes, and the reliability averages .90.

4. Space Relations is composed of a visual pattern, a portion of which is shaded. From a row of five figures the subject must decide which one or ones can be constructed from the pattern. The time to administer is 30 minutes, and the average reliability is .93.

5. Mechanical Reasoning uses 68 pairs of drawings to illustrate the mechanical principles involved in the pulley, cogs and geers, stresses and strains, transmission of power, etc. The time to administer is 30 minutes, and the reliability is .81 to .86.

6. Clerical Speed and Accuracy uses pairs of letters (small and capitals) and numbers for the test items. One of these is underlined. The problem is to find among five similar pairs that particular pair which was underlined. Time is 6 minutes. The reliability averages .87.

7. Language Usage consists of (a) spelling, and (b) sentences. Spelling consists of 100 words, some of which are misspelled, to be marked R if right, W if wrong. Time is 10 minutes and the reliability is .92. Next there are 50 sentences, each divided by lines into five parts lettered A, B, C, D, and E. Errors of grammar, punctuation, or spelling are to be recognized in the parts. It takes 25 minutes to administer, and its average reliability is .88.

¹ Psychological Corporation, New York, 1947.

These eight tests have been and are being studied. Correlations have been computed with average school marks, marks of separate subjects, intelligence tests, and other tests which purport to measure the same abilities. Thus far the tests have shown themselves to be equal to and in many cases superior to other tests in this field.

SUMMARY

Tests for two areas of the fine arts have been considered: (1) for music, and (2) for art. In both these areas, tests for capacity and tests for achievement and appreciation have been introduced. In each case, there have been attempts to analyze the larger area into its fundamental characteristics. In both cases, the combination of elements did not make the whole. In music, there seemed to be other factors in addition to pitch, rhythm, timbre, intensity, time, and memory. In art, color, line, proportion, perspective, and memory did not constitute the whole of an art object, although efficiency in these characteristics was indicative of good aptitude in art.

Tests of achievement in music and in art were not as well developed as tests of aptitude. Real achievement in both music and art consists of products which have to be rated. Aspects of musical achievement are measured in sight singing and recognition of tunes from the written notes. In art, achievement may be measured by the ability to copy a design, draw a described man, or construct a cartoon.

The reliability of these tests is generally satisfactory. Their validity is always in doubt because of the lack of an indisputable criterion of achievement. If, for example, we correlate the Seashore music test with school marks in a music course we are using a criterion which is composed of music, class attendance, and intelligence. The criterion of marks in art courses also is a mixture, so that when used it gives no certain indication of the presence of the characteristic we wish to measure.

When we turn to manual arts we find a similar story. We do have good tests of aptitudes. None of the tests, however, cover adequately the objectives set down as outcomes of instruction. Test of manual arts are divided into (1) tests of technical and related information, and (2) tests of actual performance. Actual performance may, for example, be measured by the efficiency with which a piece of wood is fashioned into an object described in a working drawing.

The tests of mechanical aptitude and ability are divided also into (1) tests of information about mechanical ability, and (2) mechanical assembly or performance tests. In selecting tests of information great care must be exercised to avoid getting so-called tests of mechanical ability which correlate highly with intelligence. Our best test in this

area is the Minnesota Mechanical Assembly Test because of the manner in which it was constructed. The builders of this test took the trouble to establish a criterion of success which could be depended upon. Once this criterion was established they could check the items of their test against it and thus ensure their efficiency. There are few satisfactory tests in the field of home economics although rating scales and check lists are available.

LIST OF TESTS OF MUSIC, ART, HOME ECONOMICS, AND MECHANICAL ABILITY

I. MUSICAL APTITUDE

1. Seashore Measures of Musical Talent, revised edition, grades 5-16 and adults. 1919-1939. Two series of three records each. Series A, for the testing of unselected groups in general surveys; Series B, for the testing of musicians and prospective or actual students of music. Blanks on which to record judgments. Time: 60-80 minutes. Authors: Carl E. Seashore, Don Lewis, and Joseph S. Saetveit. R.C.A. Manufacturing Company, Inc., Camden, N.J.

2. Kwalwasser-Dykema Music Tests, grades 4-16 and adults. 1930. One form. Time: 60 minutes. Authors: Jakob Kwalwasser and Peter W. Dykema. Carl Fischer, Inc., New York.

3. Drake Musical Memory Test, A Test of Musical Talent, ages 8 and over. 1934. Two forms. Time: 25 minutes. Author: Raleigh M. Drake. Public School Publishing Company, Bloomington, Ill.

4. Musical Aptitude Test, Series A, grades 4-10. 1950. Tests given with piano. Time: 40-50 minutes. Authors: Harvey S. Whistler and Louis P. Thorpe. California Test Bureau, Los Angeles, Calif.

II. MUSICAL ACHIEVEMENT

1. Beach Music Test, grades 4-16. 1920-1939. One form. Time: 40 minutes. Authors: Frank A. Beach and H. E. Schrammel. Kansas State Teachers College, Emporia, Kans.

2. Knuth Achievement Tests in Music, grades 3-12. 1936. Two forms.

Three levels. Division *a*, grades 3-4; Division *b*, grades 5-6; Division *c*, grades 7-12. Nontimed (40-45 minutes). Author: William E. Knuth. Educational Testing Bureau, Minneapolis.

3. Strause Music Test, grades 4-16. 1937. Three forms. Time: 60 minutes. A general achievement test. Authors: Catherine E. Strause and H. E. Schrammel. Kansas State Teachers College, Emporia, Kans.

4. Kwalwasser-Ruch Tests of Musical Accomplishment, grades 4-12. 1924-1927. Ten parts. Authors: Jacob Kwalwasser and G. M. Ruch. Bureau of Educational Research and Service, University of Iowa, Iowa City. Time: 40-50 minutes.

5. Kwalwasser Test of Musical Information and Appreciation, grades 9-16. 1927. One form. Time: 40 minutes. Author: Jacob Kwalwasser. Bureau of Educational Research and Service, University of Iowa, Iowa City.

III. ART

1. Horn Art Aptitude Inventory, preliminary form, 1944 revision, grades 12-16. One form. Time: 50 minutes. Author: Charles C. Horn. Office of Educational Research, Rochester Institute of Technology, Rochester, N.Y.

2. Meier-Seashore Art Judgment Test, grades 7-12. 1929-1930. One form (125 paired pictures). Time: 45-50 minutes. Authors: Norman Charles Meier and Carl Emil Seashore (see text). Bureau of Educational Research and Service, State University of Iowa, Iowa City.

3. Meier Art Test, I—Art Judgment Test. One form (100 paired pictures). Nontimed (45–60 minutes). Author: Norman Charles Meier. Bureau of Educational Research and Service, University of Iowa, Iowa City.

4. McAdory Art Test, all grades, colleges, and art schools. 1929. One form, a folio of 72 plates. Nontimed (about 90 minutes). Author: Margaret McAdory (see text). Bureau of Publications, Teachers College, Columbia University, New York.

5. Tests in Fundamental Abilities of Visual Arts, grades 3–12. 1927. One form. Three parts. Time: 30 (35 minutes). Author: Alfred S. Lewerenz (see text). California Test Bureau, Los Angeles, Calif.

6. Knauber Art Ability Test, grades 7–16 and adults. 1932–1935. One form. Nontimed (180 minutes). Author: Alma Jordan Knauber (see text). Published by the author, Cincinnati, Ohio.

IV. HOME ECONOMICS

1. Engle-Stenquist Home Economics Test, grades 5–10. 1931. Two forms, A and B. Time: 60 minutes. Authors: Edna M. Engle and John L. Stenquist. World Book Company Yonkers, N.Y. (out of print).

2. State High School Tests for Indiana, grades 7–8. 1945–1946. Four tests: (1) assisting with care and play of children, (2) assisting with clothing problems, (3) helping with food in the home, and (4) helping with the housekeeping. Time: 28 minutes for each test. Authors: Test 1, Alice Stair and Muriel G. McFarland; Test 2, Elizabeth Anderson, Muriel G. McFarland, and Kathleen McGillicuddy; Test 3, Elizabeth Anderson and Muriel G. McFarland; Test 4, Evelyn Swaim, Kathleen McGillicuddy, and Muriel G. McFarland. State High School Testing Service, Purdue University, Lafayette, Ind.

3. State High School Tests for Indiana high school. 1943–1947. Seven tests: (1) child development, (2) clothing I,

(3) clothing II, (4) foods I, food selection and preparation, (5) foods II, planning for family food needs, (6) home care of the sick, (7) housing the family. Time: 55–60 minutes. Authors: Test 1, Roberta Kelly, Alice Stair, and Muriel G. McFarland; Test 2, Mary I. Healey, Jeannette O. Parvis, and Muriel G. McFarland; Test 3, Mary I. Healey, Ruth Davis Moutoux, Jeannette O. Parvis, Louise Stedman, and Muriel G. McFarland; Tests 4 and 5, Mary T. Swickard and Muriel G. McFarland; Test 6, Jeannette O. Parvis, Gleela Ratcliffe, Ruth Davis, and Muriel G. McFarland; Test 7, Jeannette O. Parvis and Muriel G. McFarland. State High School Testing Service, Purdue University, Lafayette, Ind.

4. Minnesota Check List for Food Preparation and Serving, revised edition, grades 7–16. 1945. Author: Clara M. Brown. University of Minnesota Press Minneapolis.

5. *Minnesota Food Score Cards*, high school and college. 1946. Author: Clara M. Brown. Cooperative Test Service, New York.

6. Unit Scales of Attainment in Foods and Household Management, grades 7–9. 1933. Two forms. Nontimed (50 minutes). Authors: Ethel B. Reeve and Clara M. Brown. Educational Test Bureau, Minneapolis.

7. Tests in Comprehension of Patterns, grades 6–12. 1927. One form. Nontimed. Authors: L. Stevenson and M. Trilling. Public School Publishing Company, Bloomington, Ill.

V. MECHANICAL ABILITY

1. O'Connor Finger Dexterity Test, 13 years and above. 310 metal pegs or pins, 1 inch in length; a metal plate with 100 holes, each hole large enough for three pins; pins picked up with the fingers three at the time and placed in each hole until all holes are filled. Time: 8–10 minutes. Stevens Institute of Technology, Hoboken, N.J.

2. O'Connor Tweezer Dexterity Test, about 13 years and above. 100 metal pins as above; subject picks up one pin at the time with small tweezers and places one pin in each hole. Time: 8-10 minutes. Stevens Institute of Technology, Hoboken, N.J.

3. Minnesota Manual Dexterity Test, 13 years and above. Consists of four rows of 15 blocks each. Score is the time it takes (1) to pick up the blocks with one hand and put them in the hole, or (2) to pick them up with one hand turn them over with the other and put them back, or (3) to move each block to next hole above. Test of speed. Author: W. Z. Ziegler. University of Minnesota, Minneapolis.

4. I.E.R. Assembly Test for Girls, shortened form. Originally constructed by H. A. Toops, and shortened by Emily T. Burr and Zaida M. Metcalf. Time: 25-30 minutes. Norms adopted by Burr and Metcalf from experience. C. H. Stoelting, Chicago, Ill.

5. O'Rourke Mechanical Aptitude Tests, ages 15-24. Two parts. In Part I the problem is to select which of several tools would be used with certain pictured objects. Part II is entirely verbal and consists of 60 questions of a mechanical nature presented in a multiple-choice form. Time: Part I, 30 minutes; Part II, 25 minutes. Psychological Corporation, New York.

6. Stenquist Mechanical Aptitude Tests I and II, boys aged 12-15. Test I is made up of 95 problems which consist of finding out which of five pictures belongs with one of five other pictures. Test II, which is somewhat like Test I, contains also some diagrams of machine parts. World Book Company, Yonkers, N.Y.

7. Revised Minnesota Paper Form Board (see text), boys aged 9 and over, and men. Authors: Rensis Likert and William H. Quasha. Psychological Corporation, New York.

8. MacQuarrie Test for Mechanical Ability (see text). Author: T. W. MacQuarrie. California Test Bureau, Los Angeles Calif.

9. Minnesota Mechanical Assembly Test, junior and senior high school and men (see text). Authors: D. G. Paterson, R. M. Elliot, L. D. Anderson, and Edna Heidbreder. Marietta Apparatus Co., Marietta, Ohio.

10. Prognostic Test of Mechanical Abilities, grade 7 to adult. 1950. Time: 45 minutes. Authors: J. Wayne Wrightstone and Charles E. O'Toole. California Test Bureau, Los Angeles, Calif.

11. Minnesota Spatial Relations Test, upper elementary grades, high school, and adults. Consists of four standard form boards, A, B, C, D. From each form board, 58 pieces differing in form and size are cut. Time to put all pieces back into boards B, C, and D is the score (board A is used for practice). Time: 15-45 minutes. Authors: Donald G. Paterson, Richard M. Elliott, L. Dewey Anderson, H. A. Toops, and Edna Heidbreder. Marietta Apparatus Company, Marietta, Ohio.

12. Mellenbruch Mechanical Aptitude Test for Men and Women (see text), grades 7-16 and adults. Author: P. L. Mellenbruch. Science Research Associates, Chicago, Ill.

13. Test of Mechanical Comprehension, grade 9 and over. Authors: George K. Bennett and Dinah E. Frye (see text). Psychological Corporation, New York.

QUESTIONS AND EXERCISES

1. Describe the main features of Seashore's Measures of Musical Talents. What success has it had as a predictor of musical accomplishment?

2. What other factors enter into musical accomplishment in addition to those included in the Seashore tests?

3. Explain the difficulties which en-

ter into the measurement of musical achievement.

4. Summarize the uses of tests in music.

5. What are the salient features of the Meier-Seashore Art Judgment Test? Compare it in detail with the McAdory Art Test. What are two weaknesses of the latter test?

6. Do you agree that the Knauber Art Ability Test is well named?

7. How are norms of achievement tests in fine arts established?

8. What is the correct procedure in relation to manual arts when students

are discovered with a proved inadequacy in academic subjects?

9. Describe the procedure used (a) to construct the Newkirk-Stoddard Home Mechanics Test, and (b) to establish the criterion for the Minnesota Mechanical Assembly Test. Why is this latter procedure so highly regarded?

10. How is the predictive capacity of a test indicated? How efficient is a prediction based on a correlation of .60?

11. What explanation might be advanced for including the measurement of music, art, home economics, and mechanical aptitude in one chapter?

BIBLIOGRAPHY

I. MUSIC

DRAKE, RALEIGH M.: "The Validity and Reliability of Tests of Musical Talent," *Journal of Applied Psychology* (1933) 17:447-458.

FARNSWORTH, PAUL R.: "Are 'Music Capacity' Tests More Important than 'Intelligence Tests' in the Prediction of Several Types of Musical Grades?" *Journal Applied Psychology* (1935) 19: 347-350.

GREENE, EDWARD B.: *Measurements of Human Behavior*, pp. 425-438. New York: The Odyssey Press, Inc., 1941.

HIGHSMITH, J. A.: "Selecting Musical Talent," *Journal of Applied Psychology* (1929) 13:486-493.

JOHNSON, GUY B.: "A Summary of Negro Scores on the Seashore Musical Talent Tests," *Journal of Comparative Psychology* (1931) 11:383-393.

KNUTH, WILLIAM E.: *The Construction and Validation of Music Tests Designed to Measure Certain Aspects of Sight Reading*, unpublished doctor's thesis, University of California, 1932.

MURSELL, JAMES L.: *The Psychology of Music*. New York: W. W. Norton & Company, 1937.

Predicting Success in the Study of Music, Veterans Administration Technical Bulletin TB7-77, Dec. 31, 1947.

SCHOEN, MAX: *The Psychology of Music: A Survey for Teacher and Musi-*

cian. New York: The Ronald Press Company, 1940.

SEASHORE, CARL E.: *Psychology of Music*. New York: McGraw-Hill Book Company, Inc., 1938.

———: *In Search of Beauty in Music*. New York: The Ronald Press Company, 1947.

STANTON, HAZEL M.: *Prognosis of Musical Achievement*. Rochester, N.Y.: Eastman School of Music, University of Rochester, 1929.

II. ART

CARROLL, HERBERT A.: "What Do the Meier-Seashore and the McAdory Art Tests Measure?" *Journal of Educational Research* (1933) 26:661-665.

FAULKNER, RAY: "Standards of Value in Art," "Art in American Life and Education," *Fortieth Yearbook of the National Society for the Study of Education*, Chap. XXVII, pp. 401-426. Bloomington, Ill.: Public School Publishing Company, 1941.

———: *An Experimental Investigation Designed to Develop Tests to Measure Art Understanding and Appreciation*, unpublished doctor's thesis, University of Minnesota, 1937.

GREENE, EDWARD B.: *Measurements of Human Behavior*, Chap. 13. New York: The Odyssey Press, Inc., 1941.

KINTNER, MADALINE: *The Measurement of Artistic Abilities*. New York: Psychological Corporation, 1933.

KNAUBER, ALMA JORDAN: "The Construction and Standardization of the Knauber Art Tests," *Education* (1935) 56:165-170.

LEWERENZ, ALFRED S.: "Predicting Ability in Art," *Journal of Educational Psychology* (1929) 20:702-704.

MEIER, NORMAN C.: "Recent Research in the Psychology of Art," "Art in American Life and Education," *Fortieth Yearbook of the National Society for the Study of Education*, Chap. XXVI. Bloomington, Ill.: Public School Publishing Company, 1941.

III. MANUAL ARTS

BABCOCK, HARRIET, and MARION RINES EMERSON: "An Analytical Study of the MacQuarrie Test for Mechanical Ability," *Journal of Educational Psychology*, (1938) 29:50-55.

BENNETT, GEORGE K., and RUTH M. CRUICKSHANK: "Sex Differences in the Understanding of Mechanical Problems," *Journal of Applied Psychology* (1942) 26:121-127.

——— and ———: *A Summary of Manual and Mechanical Ability Tests*. New York: Psychological Corporation, 1942.

BINGHAM, WALTER VAN DYKE: *Aptitudes and Aptitude Testing*. New York: Harper & Brothers, 1937.

DURAN, JUNE C.: *MacQuarrie Test for Mechanical Ability*. Los Angeles, Calif.: California Test Bureau.

HORNING, S. D., and RUTH S. LEONARD: "Testing Mechanical Ability by the MacQuarrie Test," *Industrial Arts Magazine* (1926) 15:348-350.

MORGAN, W. J.: "Some Remarks and Results of Aptitude Testing in Technical and Industrial Schools," *Journal of Social Psychology* (1944) 20:19-29.

NEWKIRK, LOUIS V.: *Validating and Testing Home Mechanics Content*, Studies in Education, Vol. 6, No. 4, University of Iowa, 1930-1932.

PATERSON, DONALD G., et al.: *Minnesota Mechanical Ability Tests*. Minneapolis: University of Minnesota Press, 1930.

PERRY, FAY V., and M. E. BROOM: "A Study of Standard Tests and of Teacher Made Objective Tests in Foods," *Journal of Educational Research* (1932) 26:102-104.

STOY, E. G.: "Additional Tests for Mechanical Drawing Aptitude," *Personnel Journal* (1928) 6:361-366.

TIFFIN, JOSEPH: *Industrial Psychology*, 2d ed. New York: Prentice-Hall, Inc., 1947.

CHAPTER 13

Measurement of Physical Education and Health

Studies of the physical condition and general health of the draftees in both the First World War and the Second World War have clearly shown that hundreds of thousands of our young men were in such poor physical health that they were doubtful risks as members of our armed forces. The knowledge of such conditions has brought about a renewed interest in the improvement of the physical condition and general health of all people. Particularly has this movement influenced the physical-education programs *for all* students in our schools and colleges.

As in other areas of instruction, improvement comes with a greater degree of certainty when (1) objectives are clearly defined, (2) measuring instruments which indicate progress toward the objective are provided, and (3) procedures of instruction are modified in the light of objective measures.

OBJECTIVES IN PHYSICAL EDUCATION

That objectives in instruction in physical education reflect the best present philosophy in education is indicated by such lists as have been prepared by its teachers. Many leaders in this field would agree that the development of skills (neuromuscular), physical fitness, and social efficiency constitutes the general purpose of instruction in physical education.¹ There undoubtedly would also be general agreement with La Porte's analysis of objectives.² Included in this more detailed list are:

1. The development of skills—athletic, gymnastic, aquatic, rhythmic—for immediate educational purposes as well as for use later in leisure time. This would involve also a knowledge of the rules, techniques, etc., of certain skills.
2. Development of social standards, appreciations, and attitudes by

¹ See Bovard, John F., Frederick W. Cozens, and E. Patricia Hagman, *Tests and Measurements in Physical Education*, 3d ed. p. 5. Philadelphia: W. B. Saunders Company, 1949.

² La Porte, William L., "Ten Major Objectives of Health and Physical Education," *California Physical Education Health and Recreation Journal*, January, 1936, p. 6. Permission for use from Professor William L. La Porte.

means of intensive participation in sports and games under favorable conditions of leadership.

3. Development of certain personality traits such as poise, self-confidence, and self-expression, which come as a result of having each student participate in certain activities. Such participation also results in development of leadership capacities.

4. Development of safety habits in actual life situations so that they will be continued in later life.

5. Elimination of those physical defects, such as bad posture, which are remediable.

6. Development of essential health habits, health knowledge, and health attitudes in such a way that they will function in the child's life during school and later when he becomes an adult.

From this list, only slightly modified from the original, it is clear that the aims of physical education are abundantly worthy of attainment and fit in with the improvement of the whole personality—an idea so prevalent in modern educational philosophy.

TESTS OF PHYSICAL CAPACITIES

It is very difficult to distinguish between capacity and ability, for the moment a child is born his environment begins to act and react upon his capacities, bringing about changes which could strictly be defined as abilities. What we shall mean by capacities includes those traits which have had no special systematic training more than occurs in the usual environment. Thus when we speak ordinarily of lung capacity, of motor capacity, of steadiness, or tapping, etc., we usually mean traits with no systematic training. On the other hand, when we think of basketball or tennis we think of skills or abilities. Tests of physical capacities, therefore, indicate a child's possibilities which we have to work with and develop.

CARDIOVASCULAR TESTS

These tests of pulse rate and blood pressure are basal to most kinds of physical development. The discovery of their relation to the general condition of muscular tonus was a great advance. It was found, for example, that as the body assumed an erect position the force of gravity caused in a normal person an increase in pulse rate and a momentary decrease followed by an increase in systolic blood pressure. Furthermore, it was discovered that the speed with which these two measures returned to normal indicated the efficiency of the circulatory system.

Systolic pressure and pulse rate are the two factors measured by the Schneider Test.¹ The pulse rate and systolic blood pressure are taken

¹ Schneider, E. C., "A Cardiovascular Rating as a Measure of Physical Fatigue and Efficiency," *Journal of the American Medical Association* (1920) 74:1507.

two or three times during 5 minutes of rest in a reclining position. The subject then assumes an erect position. After a delay of 2 to 3 minutes pulse rate and systolic blood pressure are taken and recorded. The difference between the readings (1) when reclining and (2) when standing are indicative of the general physical condition. A second part of this test consists of measuring pulse rate and blood pressure before and after exercise. The exercise consists of placing one's right foot in a chair 18 inches high and then bringing the left foot slowly to the side of the right one, once every 3 seconds for 15 seconds. After the exercise, the pulse rate is read at intervals of 60 seconds, 90 seconds, and 120 seconds. Tables are furnished which make the scoring easy. The total points are 18 for a perfect record. A score of 9 points or less indicates deficiency.

The Harvard Step Test, developed during the Second World War, does not bother with pretesting but uses much more intense exercise. In this test the exercise consists of stepping up on and down from, a 20-inch platform at 2-second intervals, 30 times a minute for 5 minutes, unless the individual is unable to continue before the expiration of the specified time. "Beginning exactly one minute after he stops, count the number of heartbeats for exactly 30 seconds."¹ Only two observations are necessary: (1) the duration of effort, and (2) the number of heartbeats. By means of a table it is possible to substitute these two variables and read directly an index of efficiency. For normal healthy young men this index is 50. Those men in *poor* physical condition score below 50 and those in *good* condition score above 80.

While these individual tests are undoubtedly efficient, the ordinary teacher of physical education desires a *group test* even though less precise which can be administered to 20 or 30 pupils at one time. Such a test is the Michigan Pulse Rate Test for Physical Fitness.² In this test the children are first taught to count their own pulse. After this process has been well learned they count their own pulse while standing at ease and before their exercise and make their records on the blackboard. The class then runs in place at the rate of three steps per second for 15 seconds. They must lift their feet 6 inches high at least. They again count their pulse $\frac{1}{2}$ minute after exercise, 1 minute after exercise, and at 2-minute and 3-minute intervals after exercise. They record their counts on the blackboard.

If the child's pulse returns to normal after $\frac{1}{2}$ minute his score is A; if after 1 minute, B; after 2 minutes, C; after 3 minutes, D; and E if it takes longer than 3 minutes. If his pulse is irregular his grade drops one rank.

¹ Morehouse, Lawrence E., and Augustus T. Miller, Jr., *Physiology of Exercise*, p. 274. St. Louis: The C. V. Mosby Company, Medical Publishers, 1948.

² "Physical Education in the State of Michigan," *American Physical Education Review* (1920) 25:138-139.

A second test, more inclusive but built on the same principle, is the California Group Functional Test.¹ This test may be divided into four parts:

1. In the first part the body weight is considered in its relation to age and height. Needed figures are secured from the American Child Health Association.

2. In the second part, the breath-holding test, children with faces oriented toward the blackboard hold their breath as long as possible while the leader counts aloud elapsing seconds. When each child exhales, he records the time on the blackboard.

3. The third part has the children count their pulse *before* and *after* doing 25 forward body bends in 30 seconds. The children while facing the board count their pulse for 30 seconds, then stand at ease for 90 seconds, *then* count their pulse for 30 seconds and record the count on the blackboard.

4. The records for the potato race for the girls and for the boys $\frac{1}{2}$ mile are also kept. Supplementary data are collected (1) of children who are excused from the test at their own request, (2) of children who give up during the test, (3) of children that the leader thought it best to stop during the test, (4) of children that showed marked breathlessness after the test, and (5) of those that showed marked fatigue. There is no report of the reliability or validity of these last two tests. In addition, there are errors appearing in the record because the children might not record their pulse rates correctly.

"The cardiovascular tests are of limited use to the average teacher of physical education," say Bovard, Cozens, and Hagman.² They point out that the reliability of such measures is affected by age, sex, temperature, climate, humidity, emotional conditions, and altitude.

TESTS OF STRENGTH

Along with general physical fitness as indicated by the cardiovascular tests is that of physical strength. Static strength of the hand (grip), back, and legs are well measured by a variety of dynamometers. The word *dynamometer* comes from two Greek words which mean to measure strength or power. Strength in action has been measured by dipping on parallel bars, by chinning, and by the ergograph. The ergograph keeps a record, let us say, of lifting an 8-pound weight with the middle finger each second until fatigue sets in. The weight is attached to a string which works over a pulley and is attached to the middle finger. Lung capacity is measured by the spirometer, into which an individual

¹ Stolz, H. R., "Group Functional Tests," Circular Letter M 30, Nov. 7, 1923. Sacramento, Calif.: California State Board of Education, Department of Physical Education, 1923.

² *Op. cit.*, p. 87.

breathes all the air he has previously packed into his lungs. What the physical-education teacher would like is some way of combining these different measures of strength into a simple index.

The Rogers Strength Index

The Rogers Strength Index¹ recommends itself because of its simplicity and effectiveness. The index is secured by adding the scores secured in the following manner:

1. Number of cubic inches in lung capacity²
2. Number of pounds pressure in right grip
3. Number of pounds pressure in left grip
4. Number of pounds lifted, using back
5. Number of pounds lifted using legs
6. Strength of arms (pull-ups + push-ups) $\times \left[\frac{\text{weight}}{10} + (\text{height in inches} - 60 \text{ inches}) \right]$

A physical-fitness index may be computed from the strength index by dividing the strength index by age and weight norms times 100. Its author claims that this test is a highly valid measure, that it is two and one-half times as accurate as the use of weight alone and almost twice as accurate as the optimal combination of age, height, and weight. All tests can be given at the rate of one boy per minute and indices can be computed in a few seconds. Furthermore, the tests are easily scored and interesting to take. The strength index also is highly reliable.

While measures of strength do not correlate very highly with the athletic ability of girls, one author³ offers a weighted strength index which consists of: 5 (thigh flexors) + 7 (push-ups) + 1 (leg lift) as a measure of girls' strength. This index correlates .49 with the athletic ability of girls. Finally, weighted tests have been devised for measuring the strength of junior and senior high school students. The indices are:

1. Boys' strength: .1 broad jump + 2.3 shot-put (4 pounds) + weight
2. Girls' strength: .5 broad jump + 3 shot-put + weight⁴

¹ Rogers, Frederick Rand, *Tests and Measurements Programs in the Redirection of Physical Education*. New York: Bureau of Publications, Teachers College, Columbia University, 1927.

² Rogers, Frederick Rand, *Physical Capacity Tests in the Administration of Physical Education*. New York: Bureau of Publications, Teachers College, Columbia University, 1925.

³ Anderson, Theresa W., "Weighted Strength Tests for the Prediction of Athletic Ability in High School Girls," *Research Quarterly of the American Association for Health, Physical Education and Recreation* (1936) 7:136-142.

⁴ Stansbury, Edgar, "A Simplified Method of Classifying Junior and Senior Boys into Homogeneous Groups for Physical Education Activities," *Research Quarterly of the American Association for Health, Physical Education and Recreation* (1941) 12:765-776. •

TESTS OF POSTURE

Good posture is more of a condition than a capacity.

The best measures of posture are obtained from photographing the subject against a board marked off in quadrilles. The subject stands on a turntable placed at a known distance from the quadrille board. Photographs are made of the individual from different positions and measured results can be secured quickly and accurately. Unfortunately most schools are not equipped with cameras, dark rooms, and quadrille

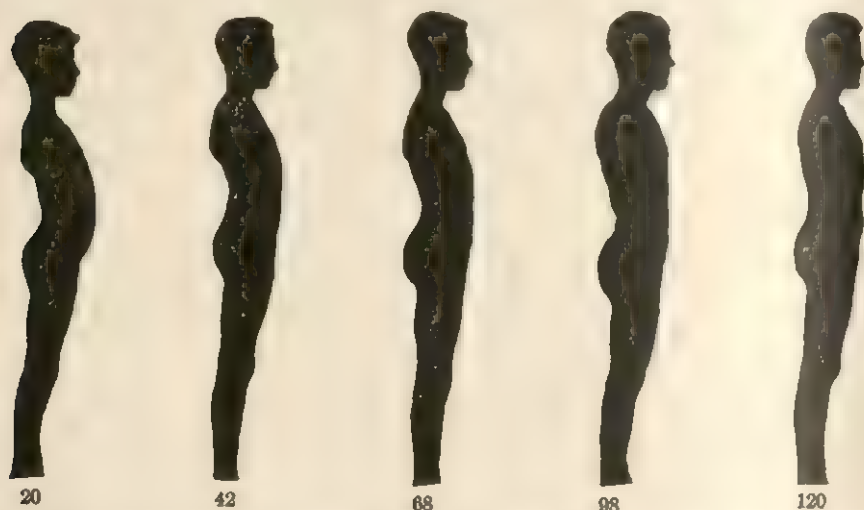


FIG. 28. Samples of silhouette scale (Clifford L. Brownell). (By permission of Bureau of Publications, Teachers College, Columbia University, New York.)

boards, hence posture becomes more a matter of rating than of exact measurement.¹

Ordinary observation, however, can be greatly improved by the use of rating scales which contain silhouettes of postures of increasing goodness. Such a scale is the Brownell scale (Fig. 28) for measuring anterior-posterior posture.² This author gathered 100 silhouettes randomly and had them arranged in order of merit by a group of experts. From this 100, 13 samples were selected and arranged in a scale. Under each silhouette is placed the scale score whose value was statistically determined. This scale may be used as was the handwriting scale of Thorndike. One simply moves the silhouette of a child up the scale until the

¹ See Bovard, Cozens, and Hagman, *op. cit.*, pp. 42-45.

² Brownell, Clifford L., *A Scale for Measuring the Anterior-Posterior Posture of Ninth Grade Boys*. Contributions to Education, No. 325. New York: Bureau of Publications, Teachers College, Columbia University, 1928.

next one seems better and then starting at the top moves the sample down until the next one seems worse. The average of the two scores thus secured is the child's posture score.

TESTS OF MOTOR COORDINATION

The last of these capacities to be considered here is that of motor coordination. How do quickness of reaction, strength, breathing, etc., work together in performance? It is this integration of action directed toward a certain goal that we think of under the term "motor coordination."

*Brace Scale of Motor Ability Tests*¹

This scale or set of tests is made up of two batteries of 10 events each, which are easy to give and to score. It is suitable for ages 8 to 18. The following samples indicate the nature of the tests:

1. Walking in a straight line, heel to toe, for ten steps
4. Kneel on both knees, with arms folded behind the back and stand
7. Full turn left in the air and land without losing balance
10. Jumping through a loop formed by grasping one toe with opposite hand
13. Bend forward, place both hands on the floor, raise the right leg, touch forehead to the floor, and stand without losing balance
16. Jumping to feet from kneeling position
19. Frog stand for 5 seconds
20. One knee dip with foot extended forward and recover position

There is also an Iowa Revision of the Brace Scale of Motor Ability Tests.² McCloy, who did the revision, tried out 40 stunts and eliminated them one by one until he had 21 items left. He retained 10 of the items of the original battery, added some new material and modified the administration and scoring procedures. In his textbook McCloy gives detailed instructions for administering and scoring the test and for giving the test to groups of subjects. He claimed that these changes improved the validity of the test. In the original test Brace thought that such measures of motor capacity would aid greatly in classifying pupils for physical education. The results of the tests do aid us in the study of special performance disabilities as well in the equating of groups in physical education.

¹ Brace, David K., *Measuring Motor Ability*. New York: A. S. Barnes and Company, 1927. Items by permission.

² McCloy, C. H., *Tests and Measurements in Health and Physical Education*, pp. 70-77. New York: Appleton-Century-Crofts, Inc., 1942.

ACHIEVEMENT TESTS

Achievement tests in physical education follow the same standards of construction as do the tests we have described thus far. Their items must be carefully selected so as to be representative of the total skill or ability. The test must be sufficiently reliable. It also must be valid. The criteria against which the test is validated may be (1) scores obtained by the ratings of experts, (2) T-scores obtained from a rich variety of tests of the ability in question, and (3) scores from a round robin in which each player becomes an opponent of every other one. When possible the tests should be applicable to groups. One criterion emphasized is somewhat different from other tests. When possible it is better to have a test which may be used both as a practice test and as an indicator of achievement.¹ As in other areas, norms should be computed from representative populations.

Achievement scales in physical education have been prepared for boys and girls in elementary, junior high, and senior high schools. These tests along with their instructions for administering and scoring appear in three volumes.² Let us consider first Achievement Scales in Physical Education Activities, which also includes in its title "for Boys and Girls in Elementary and Junior High Schools." In this book instructions for administering 33 different activities are carefully described and T-score norms are furnished for eight different classifications from A to H. Classifications of children are based on a table of standards of (1) height in inches, (2) age in years and months, and (3) weight in pounds. Suppose we had a child who is 54 inches tall, is 12 years and 7 months old, and weighs 104 pounds. By referring to a table in this test³ we find that (1) for a height of 54 inches he receives an exponent of 4, (2) for 12-7 in age he receives an exponent of 6, and (3) for a weight of 104 pounds he receives an exponent of 9. If we add these three exponents together we get a total of 19. A total of 19 places him in Class C. By using the tables of performance we can discover the sort of record this child has in comparison with others who are classified as Class C. It is thus seen that a child is compared only with those in *his* class. Samples of the 33 items are basketball throw for distance, jump and reach, playground baseball,

¹ *Ibid.*, pp. 169-172.

² Neilson, N. P., and Frederick W. Cozens, *Achievement Scales in Physical Education Activities*. New York: A. S. Barnes and Company, 1939. Cozens, Frederick W., Martin H. Trieb, and N. P. Neilson, *Physical Education Achievement Scales for Boys in Secondary Schools*. New York: A. S. Barnes and Company, 1936. Cozens, F. W., Hazel J. Cubberley, and N. P. Neilson, *Achievement Scales in Physical Education Activities*. New York: A. S. Barnes and Company, 1937.

³ Neilson and Cozens, *op. cit.*, p. 6.

throw for accuracy, push-up, running high jump, and standing hop, step, and jump.

The norms for classes from A to H were computed from some 79,000 children, and the scores from each event are transmuted into standard scores. In like manner, achievement scores are furnished for high school boys and for high school and college girls. Since it was shown that height, weight, and age are uncorrelated with athletic abilities after age 16, it was necessary to have only one set of scores instead of the eight in the series of tests just described. Most other achievement tests in physical education are constructed after the manner of those described here.

Achievement tests in the sports at the senior high school and college levels such as basketball, soccer, football, baseball, and tennis have not been so successful for men. In these areas judgment by experts resulting in ratings gets the best results. On the other hand, two authors have developed practical tests of considerable promise for girls.¹ In their practical manual they describe acceptable tests for badminton, basketball, field hockey, soccer, softball, speedball, tennis, and volley ball.

MEASUREMENT AND HEALTH INFORMATION

Good health is indicated, in the final analysis, by the existent physical condition at the present time. Has the subject any disease? Are the organs of his body working as they should? What of his eyes, ears, nose, and throat? Does he have the normal amount of energy for his age? The daily observation of children by a teacher who knows some of the major symptoms of disease and his referral of cases to nurse and physician are of the first importance. Another aspect of the problem relates to the prevention of poor health by practicing those habits and taking those precautions which in general lead to or continue good health.

There are two phases of this latter problem: (1) health knowledge, and (2) health practices. Unfortunately, health practices depend upon both the knowledge and the attitudes of subjects.

There is indeed no assurance that the knowledge of good health practices will lead to good health habits. The best instruction at the present time emphasizes both knowing and doing. Tests of health information are easier to construct and more certain in their results than inventories of health practices. To test the latter the good will of the subjects must be obtained so that they will report the habits that they actually practice and not those which they think the tester would like for them to practice.

¹ Scott, M. Gladys, and Esther French, *Better Teaching through Testing*. New York: A. S. Barnes and Company, 1945.

The Gates-Strang Health Knowledge Tests¹ are divided into (1) elementary tests for grades 3 to 8, and (2) advanced tests for grades 7 to 12. There are three forms for each division. These tests have been on the market since 1925 and were revised in 1937. "The items selected are based on extensive curriculum research involving an analysis of mortality, morbidity, and accident statistics, popular health sources, interests and needs of children, of different ages, and courses of study and textbooks."² The elementary tests are made up of 60 multiple-choice items which represent a rich variety of information. Such items are included as the harmfulness of bacteria, how to keep mosquitoes from growing in ponds, how tuberculosis is spread, the effect of the proper handling of garbage and sewage, etc. Two samples are:

22. The best lunch to choose in the school lunchroom is
 - a. Roast pork, bread, apple sauce..... a
 - b. Vegetable soup, baked potato, milk, cup custard b
 - c. Ice cream and chocolate cake..... c
 - d. Meat, potatoes, pie, milk..... d
 - e. Vegetable salad, crackers, iced tea..... e
43. The best way to study about the shape and size of bacteria is by watching them
 - a. Under a bright light..... a
 - b. With the naked eye..... b
 - c. In a darkened room..... c
 - d. Under a microscope..... d
 - e. Under a hand magnifying glass..... e

The advanced tests are composed entirely of 60 multiple-choice items which are more complicated than those of the elementary series. In these tests the emphasis is upon two major fields: (1) food and nutrition, and (2) the prevention and treatment of diseases. More than 35 of the 60 items are in some way related to these two headings. There are a few items on the functioning of certain organs, on the effects of alcohol and tobacco on growth, and on the best forms of exercise. Two samples are:

24. Vitamins are especially necessary for
 - a. Giving power to work and play..... a
 - b. Giving flavor to food..... b
 - c. Increasing health and growth..... c
 - d. Regulating body temperature..... d
 - e. Preventing typhoid fever..... e

¹ Items by permission of Bureau of Publications, Teachers College, Columbia University, New York.

² *Manual*, p. 1; also Gates, A. I., and Ruth Strang, "A Test in Health Knowledge," *Teachers College Record* (1925) 26:867-880. By permission.

40. We say a person is immune from a disease when
- a. He has not been near sick persons..... a
 - b. His body has made substances that protect it from the bacteria that cause the disease..... b
 - c. He has disinfected his sickroom..... c
 - d. His body resists cold and fatigue d
 - e. He has had the disease three times..... e

The reliability of these elementary and advanced tests varies from .74 to .86. Validity was determined by the selection of the items. Norms furnished consist of distributions of scores for the elementary tests secured from a large city system, from a suburban school system, and from rural schools. For the advanced tests there are score distributions obtained from a large city high school and from a suburban high school.

The tests are useful for analyzing the health knowledge of a single subject as well as for indicating the general progress of a class. They suffer somewhat from the wide variety of items tested.

The second illustration, *Health Inventory for High School Students*,¹ is distinctive for attempting to enlist the students' cooperation in securing information on their health status and practices. This inventory, suitable for grades 9 to 12, is divided into two parts: (1) health conditions, and (2) health information. This inventory is an outgrowth of several years' study of health knowledge in the city of Los Angeles. The items of the final form are based on "extensive Curriculum research involving the analysis of textbooks, courses of study, popular health sources, and other authorities on health information."²

Part I, on health conditions, is divided into (1) health status and (2) health practice. It is on this part that the cooperation of the student is enlisted.

The most common answers on the status part are "(1) Frequently (2) Occasionally (3) Never" or "(1) Frequently (2) Seldom (3) Never." Questions about being sick in bed, colds, headaches, tiredness, and toothache are asked. Two samples are:

- 3. Do you have colds?
(1) Frequently (2) Seldom (3) Never
- 8. Do your teeth hurt because of decay?
(1) Frequently (2) Occasionally (3) Never

¹ Neher, Gerwin Charles, *A Study of the Health Knowledge, Attitudes, Status and Practice of High School Pupils*, unpublished doctoral dissertation, University of Southern California, 1942.

² *Manual*, p. 1. Items by permission from California Test Bureau, Los Angeles, Calif.

There are 20 items on health practice. Questions as to whether you drink at least one pint of milk a day, maintain a correct posture, have formed a habit of daily bowel action, avoid colds and other communicable disease, or the average number of hours you sleep per night are asked. Two samples are:

13. Do you ever eat candy or other sweets just before meals?
(1) Frequently (2) Occasionally (3) Never
22. Do you use drugs such as aspirin, bromides, etc. for cure of headaches?
(1) Frequently (2) Occasionally (3) Never

The score of this part is a weighted one. If the answer selected is the perfect one the subject receives 3 points, 2 for a poorer answer, and 1 for the poorest answer. The total of these weighted points makes up the score.

Part II of this inventory consists of 69 items entitled "What You Know about Health." The subdivisions are:

1. Public health. This section asks for the definition of slum areas, the reliability of radio advertising, and the effects upon health of venereal diseases—eight questions altogether.

2. First aid. Here the test inquires about what to do if you feel faint, what to do if you have a turned ankle, how to neutralize acid spilt on the skin or clotting, etc.—seven items.

43. After sending for a physician the first thing to do for a person who has swallowed poison is to

1. Give him artificial respiration
2. Make him vomit
3. Go to the druggist for an antidote
4. Put him to bed
5. Give him a strong laxative.

43

3. Prevention of disease (15 items). This section includes questions about the pasteurization of milk, what a communicable disease is, why milk turns sour, and how best to control smallpox and diphtheria.

60. Measles is most contagious

1. Before the rash appears
2. When the rash is most noticeable
3. When the skin begins to peel
4. After the skin has peeled
5. When the rash is disappearing.

60

4. Proper health habits (12 items). Here are such questions as why breathing through the nose is best for health, what the correct amount of sleep is for high school boys and girls, and what type of bath is best when you are tired and nervous.

5. Diet (18 items). This section raises such questions as to the foods which contain the most minerals, the main food value of meat, the prevention of constipation, and how well pork should be cooked.

6. Mental hygiene (nine items). This deals with such problems as the influence of worry on health, the relation between facing life squarely and mental health, as well as the relation between poise and emotional balance.

The reliability of the test as a whole is .86. When one breaks down the 99 items into eight different parts, as is recommended in making a profile, one wonders what the reliability of each part might be. The profile, though, does help to tell at a glance just where the student is weak. If this is followed by an item analysis of the weak part, *real* diagnosis of difficulties may be attained. The norms are based on returns from 2,415 students in the city of Los Angeles and are reported in both percentile ranks and descriptive words such as very low, low, average, high, and very high.

Use for both the Gates-Strang Health Knowledge Tests and the Neher Health Inventory for High School Students would be in (1) studying with pupils or students their weaknesses in health information, (2) influencing the teaching procedures of the teachers, and (3) improving courses of study.

LIST OF TESTS OF HEALTH EDUCATION

1. Gates-Strang Health Knowledge Tests, grades 3-12. 1937. Two levels. Three forms each level. Elementary tests, grades 3-8, 40-45 minutes; advanced tests, grades 7-12, 30-35 minutes. Authors: A. I. Gates and Ruth Strang. Bureau of Publications, Teachers College, Columbia University, New York.

2. Health Inventory for High School Students, grades 9-12. 1942. Two editions. Nontimed (about 60 minutes). Author: Gerwin Neher. California Test Bureau, Los Angeles, Calif.

3. Byrd Health Attitude Scale, grades 10-14. 1940-1941. One form. Nontimed (about 35 minutes). Author: Oliver E. Byrd. Stanford University Press, Stanford University, Calif.

4. Health and Safety Education Test, State High School Tests for Indiana, high school, first and second semesters. 1946-1947, 1945-1946. Forms A and N. Time: 40-45 minutes. Authors:

Shelby Gallien and Hilda Schwehn. State High School Testing Service, Purdue University, Lafayette, Ind.

5. Health Education Test: Knowledge and Application, grades 7-16. 1946-1947. Form A. Time: 40-45 minutes. Authors: Clifford L. Brownell, John H. Shaw, and Maurice Troyer. Acorn Publishing Company, Rockville Center, N.Y.

6. Health Practice Inventory, grades 7-14. 1943. One form. Nontimed (15-29 minutes). Author: Ned B. Johns. Stanford University Press, Stanford University, Calif.

7. Trusler-Arnett Health Knowledge Test, grades 9-16. Forms A and B. Time: 50-55 minutes. Authors: V. T. Trusler, C. E. Arnett, and H. E. Schrammel. Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kan.

8. Indiana Motor Fitness Index, boys and men, grades 10-16. 1943. 60 tests.

There are 20 items on health practice. Questions as to whether you drink at least one pint of milk a day, maintain a correct posture, have formed a habit of daily bowel action, avoid colds and other communicable disease, or the average number of hours you sleep per night are asked. Two samples are:

13. Do you ever eat candy or other sweets just before meals?
(1) Frequently (2) Occasionally (3) Never
22. Do you use drugs such as aspirin, bromides, etc. for cure of headaches?
(1) Frequently (2) Occasionally (3) Never

The score of this part is a weighted one. If the answer selected is the perfect one the subject receives 3 points, 2 for a poorer answer, and 1 for the poorest answer. The total of these weighted points makes up the score.

Part II of this inventory consists of 69 items entitled "What You Know about Health." The subdivisions are:

1. Public health. This section asks for the definition of slum areas, the reliability of radio advertising, and the effects upon health of venereal diseases—eight questions altogether.

2. First aid. Here the test inquires about what to do if you feel faint, what to do if you have a turned ankle, how to neutralize acid spilt on the skin or clotting, etc.—seven items.

43. After sending for a physician the first thing to do for a person who has swallowed poison is to

1. Give him artificial respiration
2. Make him vomit
3. Go to the druggist for an antidote
4. Put him to bed
5. Give him a strong laxative.

43

3. Prevention of disease (15 items). This section includes questions about the pasteurization of milk, what a communicable disease is, why milk turns sour, and how best to control smallpox and diphtheria.

60. Measles is most contagious

1. Before the rash appears
2. When the rash is most noticeable
3. When the skin begins to peel
4. After the skin has peeled
5. When the rash is disappearing.

60

4. Proper health habits (12 items). Here are such questions as why breathing through the nose is best for health, what the correct amount of sleep is for high school boys and girls, and what type of bath is best when you are tired and nervous.

5. Diet (18 items). This section raises such questions as to the foods which contain the most minerals, the main food value of meat, the prevention of constipation, and how well pork should be cooked.

6. Mental hygiene (nine items). This deals with such problems as the influence of worry on health, the relation between facing life squarely and mental health, as well as the relation between poise and emotional balance.

The reliability of the test as a whole is .86. When one breaks down the 99 items into eight different parts, as is recommended in making a profile, one wonders what the reliability of each part might be. The profile, though, does help to tell at a glance just where the student is weak. If this is followed by an item analysis of the weak part, *real* diagnosis of difficulties may be attained. The norms are based on returns from 2,415 students in the city of Los Angeles and are reported in both percentile ranks and descriptive words such as very low, low, average, high, and very high.

Use for both the Gates-Strang Health Knowledge Tests and the Neher Health Inventory for High School Students would be in (1) studying with pupils or students their weaknesses in health information, (2) influencing the teaching procedures of the teachers, and (3) improving courses of study.

LIST OF TESTS OF HEALTH EDUCATION

1. Gates-Strang Health Knowledge Tests, grades 3-12. 1937. Two levels. Three forms each level. Elementary tests, grades 3-8, 40-45 minutes; advanced tests, grades 7-12, 30-35 minutes. Authors: A. I. Gates and Ruth Strang. Bureau of Publications, Teachers College, Columbia University, New York.

2. Health Inventory for High School Students, grades 9-12. 1942. Two editions. Nontimed (about 60 minutes). Author: Gerwin Neher. California Test Bureau, Los Angeles, Calif.

3. Byrd Health Attitude Scale, grades 10-14. 1940-1941. One form. Nontimed (about 35 minutes). Author: Oliver E. Byrd. Stanford University Press, Stanford University, Calif.

4. Health and Safety Education Test, State High School Tests for Indiana, high school, first and second semesters. 1946-1947, 1945-1946. Forms A and N. Time: 40-45 minutes. Authors:

Shelby Gallien and Hilda Schwehn. State High School Testing Service, Purdue University, Lafayette, Ind.

5. Health Education Test: Knowledge and Application, grades 7-16. 1946-1947. Form A. Time: 40-45 minutes. Authors: Clifford L. Brownell, John H. Shaw, and Maurice Troyer. Acorn Publishing Company, Rockville Center, N.Y.

6. Health Practice Inventory, grades 7-14. 1943. One form. Nontimed (15-29 minutes). Author: Ned B. Johns. Stanford University Press, Stanford University, Calif.

7. Trusler-Arnett Health Knowledge Test, grades 9-16. Forms A and B. Time: 50-55 minutes. Authors: V. T. Trusler, C. E. Arnett, and H. E. Schrammel. Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kan.

8. Indiana Motor Fitness Index, boys and men, grades 10-16. 1943. 60 tests.

Time: 50 minutes. Authors: Karl W. Bookwalter and Carolyn W. Bookwalter. Bureau of Cooperative Research and Field Service, School of Education, Indiana University, Bloomington, Ind.

9. Health Awareness Test, grades 4-8. 1937. One form. Time: 30-40 minutes. Authors: Raymond Franzen, Mayhew Derryberry, and William A.

McCall. Bureau of Publications, Teachers College, Columbia University, New York.

10. Health Test, grades 3-8. 1937-1938. Two forms. Nontimed (about 40 minutes). Authors: Robert K. Spur and Samuel Smith. Acorn Publishing Company, Rockville Center, N.Y.

TESTS OF INFORMATION IN PHYSICAL EDUCATION

In recent years more attention has been given to tests of information in physical education. Playing regulations, game situations, and knowledge of positions and tactics have offered materials for constructing objective tests. Information tests for basketball, baseball, soccer, and tennis have been constructed. In most cases these tests have not reached the publication stage. They most usually appear in the research quarterlies of the National Physical Education Association.

RATING SCALES

Rating scales in physical education have been quite successful in several areas. Attention has already been called to the Silhouette Scale by Brownell. Another scale, the diving scale, is in constant use for measuring excellence in diving. It has 10 different divisions. Here also the rating is weighted according to the difficulty of the dive. Thus a very difficult dive might receive a weight of 3 and a rating of 8 and score 24 points in all. These two rating scales are excellent illustrations of good measuring instruments of this type. The rater is trained in exactly the things to look for and he is on the scene when the rating occurs. There are also good rating scales for basketball, riding competition, and several other sports.

SUMMARY

Teaching objectives in physical education have been clearly defined. A great variety of tests and ratings indicate clearly whether or not these objectives have been reached. These instruments have been divided into tests of physical capacity, tests of health, and tests of achievement.

Tests of physical capacity measure those traits which have had little or no formal training. Pulse rate and blood pressure, lung capacity, strength, posture, and motor coordination are samples of test of physical capacity. These tests are carefully constructed, are usually standardized on large groups of subjects, and have satisfactory reliability. Achievement tests in physical education have been standardized for more than 33 different activities. Not only have T-score norms been furnished for these numerous activities, but each activity has T-score norms at eight

different levels of physical capacity which depend on height, weight, and age. These norms, worked out in three volumes and based on scores from 79,000 subjects, are quite satisfactory.

Measurement of health information is divided into (1) health knowledge, and (2) health practices. Tests of health knowledge are constructed much as are other tests of information. Their items are based on extensive curricular research to discover items common to all good courses of study. Because information about health practices depends so much upon the willingness of the subject to report what his practices are, objective standardized tests are difficult to construct in this area.

QUESTIONS AND EXERCISES

1. *a.* Under what conditions does improvement come in physical education with the greatest degree of certainty?

b. Show that the objectives of instruction in physical education agree with the modern philosophy of education.

2. *a.* Distinguish between physical capacity and physical ability.

b. Explain the principle involved in the cardiovascular tests. Illustrate with the Schneider Test.

3. *a.* Compare the Michigan Pulse Rate Test for Physical Fitness with the California Group Functional Test.

b. What are the chief characteristics of the Rogers Strength Index?

4. *a.* What is the best way to measure posture? Why is this procedure not used more widely?

b. Describe the procedure used in Brownell's Scale.

5. *a.* What are three stunts used in Brace's Scale of Motor Ability?

b. What modifications of the Brace scale were made in the Iowa revision?

6. *a.* Describe the process used to classify boys and girls so as to measure achievement.

b. What uses can be made of achievement tests?

7. *a.* Describe the leading characteristics of the Gates-Strang Health Knowledge Test; the Health Inventory for High School Students.

b. What characteristics of the latter recommend it for use?

8. *a.* Why have rating scales been so successful in certain areas (*e.g.*, diving) of physical education?

b. Why are tests for sports so hard to construct?

BIBLIOGRAPHY

Books

BOVARD, JOHN F., FREDERICK W. COZENS, and E. PATRICIA HAGMAN: *Tests and Measurements in Physical Education*, 3d ed., pp. 3-248. Philadelphia: W. B. Saunders Company, 1949.

BRACE, DAVID K.: *Measuring Motor Ability*. New York: A. S. Barnes and Company, 1927.

BROWNELL, CLIFFORD LEE: *A Scale for Measuring the Anterior-Posterior*

Posture of Ninth Grade Boys. New York: Bureau of Publications, Teachers College, Columbia University, 1928.

COZENS, F. W., HAZEL J. CUBBERLEY, and N. P. NEILSEN: *Achievement Scales in Physical Education Activities*. New York: A. S. Barnes and Company, 1937.

———, MARTIN A. TRIEB, and N. P. NEILSEN: *Physical Education Achievement Scales for Boys in Secondary Schools*. New York: A. S. Barnes and Company, 1936.

CURETON, THOMAS K.: *Physical Fitness Appraisal and Guidance*. St. Louis: The C. V. Mosby Company, Medical Publishers, 1947.

MCCLOY, CHARLES H.: *Tests and Measurements in Health and Physical Education*. New York: Appleton-Century-Crofts, Inc., 1942.

———: *Measurement of Athletic Power*, New York: A. S. Barnes and Company, 1932.

MOREHOUSE, LAWRENCE E., and AUGUSTUS T. MILLER, Jr.: *Physiology of Exercise*. St. Louis: The C. V. Mosby Company, Medical Publishers, 1948.

NATIONAL COLLEGIATE ATHLETIC ASSOCIATION: *The Official Swimming Guide*, "Official Rules for Swimming, Fancy Diving and Water Polo," pp. 164-189. New York: A. S. Barnes and Company, 1947.

NEILSEN, N. P., and FREDERICK W. COZENS: *Achievement Scales in Physical Education Activities*. New York: A. S. Barnes and Company, 1939.

ROGERS, FREDERICK RAND: *Physical Capacity Tests in the Administration of Physical Education*. New York: Bureau of Publications, Teachers College, Columbia University, 1925.

———: *Physical Capacity Tests*. New York: A. S. Barnes and Company, 1938.

SCHNEIDER, EDWARD C., and PETER V. KARPOVICH: *Physiology of Muscular Exercise*, 3d ed. Philadelphia: W. B. Saunders Company, 1948.

SCOTT, M. GLADYS, and ESTHER FRENCH: *Better Teaching through Testing*. New York: A. S. Barnes and Company, 1945.

Articles

BOOKWALTER, KARL W.: "A Critical Evaluation of Some of the Existing

Means of Classifying Boys for Physical Education," *Research Quarterly of the American Association for Health, Physical Education, and Recreation* (1939) 10:119-127.

COZENS, FREDERICK W.: "Physical Education Measurement," *Encyclopedia of Educational Research*, pp. 814-818. New York: The Macmillan Company, 1941.

CURETON, THOMAS K., JR., and LEONARD LARSON: "Strength as an Approach to Physical Fitness," Supplement to *Research Quarterly of the American Association for Health, Physical Education, and Recreation* (1941) 12: 391-405.

EDGREN, H. D.: "An Experiment in the Testing of Ability and Progress in Basketball," *Research Quarterly of the American Physical Education Association* (1932) 3:159-171.

ESPENSCHADE, ANNA: "Development of Motor Coordination in Boys and Girls," *Research Quarterly of the American Association for Health, Physical Education, and Recreation* (1947) 18:30-43.

FRENCH, ESTHER: "The Construction of Knowledge Tests in Selected Professional Courses in Physical Education," *Research Quarterly of the American Association for Health, Physical Education, and Recreation* (1943) 14:1945.

HOWLAND, AMY R.: "National Physical Education Standards for Girls," *Journal of Health and Physical Education* (1937) 8:223.

STRANG, RUTH: "Health Education," *Encyclopedia of Educational Research*, pp. 561-571. New York: The Macmillan Company, 1941.

PART TWO

Measurement of Intelligence

CHAPTER 14

Intelligence and Its Measurement

DEVELOPMENT OF INTELLIGENCE TESTS

Two factors influenced greatly the early movement for the measurement of intelligence. One of these early motivating influences was the interest in the study of *individual differences* which some scientists possessed. Another influence grew out of attempts to measure the *intelligence of the feeble-minded*. Treatment of the feeble-minded children had varied greatly from one time to another. At one period in history defective children were exposed on a hillside to die, at another, regarded with a sort of religious awe, and at still another, blamed directly for their condition and punished accordingly. More specifically, it was the attempt to educate these unfortunates which furnished one of the first motivating influences for measuring the amount of intelligence which they possessed. There were, then, these two streams of influence which stimulated the development of measuring instruments of intelligence: the theoretical one, which arose out of the general interest in individual differences, and the practical one, which stemmed from the educational problem of separating feeble-minded from normal children.

INTEREST IN INDIVIDUAL DIFFERENCES

Theoretical psychologists were attempting to discover how greatly individuals of about the same age differed in their reaction times, in their visual discrimination, and in their motor speed. The leading figure in this movement was James McKeen Cattell of Columbia University, who had studied under the great Wundt at Leipzig University. At Leipzig, Cattell was urged by Wundt to investigate the general principles of human nature, those mental processes which are present in all mankind. Cattell, on the other hand, became far more interested in the differences among men than in their likenesses. His first experiments were conducted to measure the differences among individuals in reaction times. The functions which he experimented with were narrow. In reaction time, for example, the quickness with which a subject could press down a lever when a light was flashed or a sound heard was the

function measured. Sometimes these experimentings did hit upon tests which have later proved useful.

INTEREST IN THE FEEBLEMINDED

The second stream having to do with the measurement of intelligence flowed out of France. Consideration for the education of the deaf and blind and above all for the feeble-minded originated in France. It was the education of this latter group under the leadership of Séguin that had its greatest influence. Séguin had instructed a small class of the feeble-minded at the Bicêtre and had shown that they had improved greatly. This work of Séguin stimulated Alfred Binet (1857-1911). He early became interested in the problem of intelligence testing and later in his life was given the job of separating the feeble-minded from the normal in the city of Paris. Binet's struggles to secure satisfactory tests of intelligence paralleled rather closely in time the attempts of American psychologists. His intelligence tests were first published in 1905, and were revised in 1908 and again in 1911. Collaborating with Binet was Thomas Simon, so that the tests were called the Binet-Simon tests.

It was the 1908 edition of the Binet-Simon tests which influenced most the development of Binet testing in the United States. The initial interest in the Binet tests came first from Dr. Henry Goddard, a psychologist at work with the problems of the feeble-minded at Vineland, N.J. He translated the 1908 edition of the Binet test, adapted it to American conditions, and then standardized the test for the first time (1911) on American children. Several other men who were working on problems relating to the education of the feeble-minded saw great possibilities in these new tests. Among these was Kuhlmann, who succeeded later in producing a test based on the Binet-Simon principles. In using the test standardized by Goddard, it soon became apparent that the tests in the earlier years were too easy for American children, while those in the later years were too difficult. This produced a queer effect, for a child who in the years 6 or 7 might appear to be above the average in intelligence would in the years 13 or 14 seem to be below the average. What was needed was a test which would test children correctly at each age so that if they were bright at one age they would be bright at a later age unless they had undergone some radical change in health or in environment.

It was Terman who went to work on this problem with so much energy, intelligence, and enthusiasm that he was able to publish the Stanford Revision of the Binet-Simon tests in 1916. So successfully was this revision constructed that it became the leading individual test in the United States and remained in this position until 1937 when Terman

and Merrill together published their own revision of the first Stanford Revision.

What are some of the characteristics of this Stanford Revision which caused it to become a leader? In the first place, Terman realized the weaknesses of the Goddard revision. Moreover, he found that the instructions for giving and scoring were sometimes not as clear as they might be, nor were tests always located at the right years. He studied and experimented with all the tests he could discover. By adding some, discarding others, and moving some up or down a year or two in age he finally got them to fit pretty well a design which he had in mind.

In that design the median mental age should correspond with the median chronological age. While he never quite achieved this goal the Stanford Revision more nearly reached it than any other test published at that time. He increased the number of tests from the original 54 to 90. One of the most useful new tests added by Terman was the vocabulary test. Instructions for giving and scoring were carefully set down. There were then six tests at each age from year III to year X. Above year X there were eight tests at year XII, six at year XIV, six at average adult, and six at superior adult. There was introduced also the notion of the I.Q., which has proved to be a very practical device in spite of the recent many misgivings about its use. The credit for developing the notion of the I.Q. is usually given to Wilhelm Stern.

INDIVIDUAL TESTS OF INTELLIGENCE

MENTAL AGE SCALES

Revision of the Stanford-Binet

The Stanford-Binet showed some weaknesses quite soon after it was first published in 1916. Some items were so difficult to score that equally competent people disagreed on the outcome. Rudolph Pintner and staff at Teachers College, for example, worked out elaborate directions and illustrations for scoring this test. Lists of definitions were made which would be acceptable for passing the words of the vocabulary test and a variety of drawings to aid in scoring the diamond and the ball and field. It was, of course, clearly evident that the test did not extend low enough to test infants or high enough for the brightest, and that it omitted tests at years 11 and 13. Then, too, a curious thing would occur to the I.Q.s of the very bright. These measures of intellectual alertness seemed to shrink as the child grew older. An I.Q. of 140 at 12 years would be nearer 125 at 15 years than 140. There was also only *one form*, which was a clear drawback in the rare cases where subjects had been coached or where, for some other reason, a test needed to be repeated. To be sure, Herring had constructed a test which he claimed could be used as

an alternate form for the Stanford-Binet, but there was no truly parallel form in which the one form was made point for point like the other.

Finally, many psychologists and educators believed that the Stanford-Binet was entirely too verbal and that measures of memory played too great a role in its scores. It was perfectly apparent, then, that another test needed to be constructed after the Binet method. This test eventually became the Terman-Merrill Revision. It is hardly necessary to more than mention the careful selection and study of the available tests, their preliminary application to subjects who had taken the Stanford-Binet previously, and to tests being tentatively assigned to that age at which 50 per cent of the subjects were successful.

The old principle of *steeply increasing percentages of correct responses from one year to the next* was retained. This means that such a test as counting 13 correctly would be passed by a much larger percentage of children at year 6 than at year 5, etc. A new feature was the use of a new criterion for selecting items. This new criterion means that no item is a good one unless more of the competent subjects pass an item than of the incompetent ones. The mean age of the subjects passing the test was computed followed by the computation of the mean age of those who failed. The difference between these mean ages divided by the standard error of the difference constituted the weight. In this manner the scores of all the subjects who took the test entered into its weight or value. It was thus that the criterion of validity was satisfied. The other criteria used for selecting tests were: (1) ease and objectivity of scoring, and (2) various practical considerations such as length of time, interest to the subject, and need for variety. Altogether 209 tests for Form L and 199 for Form M were selected for the final tryout. When all tests had been discarded which for some reason or other did not fit, 129 tests were left for each form. It took six different revisions of Form L before the authors were satisfied with their test. Once Form L was constructed each item was matched point for point in the construction of Form M. Form M, then, has the same range, difficulty, and reliability as Form L.

Probably in no case was greater care exercised than in the selection of the population for the final standardization. The authors wanted the mental ages computed from this test to represent the total population of the United States. To do this, they sampled pupils from 11 states representing the various geographical areas of our country. Not only so, but it was seen, too, that the same proportions of the various socioeconomic levels should be represented in the sample population as were present in the total population. For example, 3.1 per cent of the employed males of the United States are in the professions. The authors wished to have 3.1 per cent of the children of their sample from this group. They did succeed in getting 4.5 per cent. In the semiskilled

occupations there are 30.6 per cent of employed men in the general population, Terman and Merrill secured 31.4 per cent of children from that group for their sample. Never could they get enough children from the day laborers, and so they had to allow for this weakness by standardizing their test with an average I.Q. of 102 at each age so that the representative I.Q. of the total population might be 100.

Whatever else a test has, it must have reliability. This 1937 revision does have excellent reliability. If we say a test should have at least a reliability of .90, then this is better than that, for its reliability is represented by a coefficient of .93. Curiously enough the higher I.Q.s pull this reliability down. With feeble-minded children the reliability coefficient is .98, while with very bright ones this coefficient is only .89.¹

The following samples are taken from the Terman-Merrill Revision of the Stanford-Binet. The test begins at year II and extends through the superior adult level. By inspecting samples at 3-year intervals, the rise in the level of difficulty can be more easily sensed.

Year VI

1. Defines five out of 10 or more such words as *orange, straw, gown, roar*.
2. Copies from memory a bead chain of seven beads which are alternately square and round.
3. Discovers what parts of mutilated pictures are missing.
4. Can count 3, 9, 5, and 7 blocks correctly.
5. Can discriminate between drawings that are rather obviously different.
6. Can trace two of three rather simple maze patterns.

Year IX

1. Can draw lines to represent creases and a cut-out in a paper simply folded.
2. Can detect simple verbal absurdities.
3. Can draw Greek key pattern and truncated cone pattern from memory after having seen them for 10 seconds.
4. Can give rhymes such as the name of a color that rhymes with "head."
5. Makes change mentally when he is supposedly sent to a store with 10 cents to buy 4 cents worth of candy.
6. Repeats in reverse order four digits arranged in haphazard order.

¹ Terman, Lewis M., and Maude A. Merrill, *Measuring Intelligence*, p. 46. Boston: Houghton Mifflin Company, 1937. Items by permission.

Year XII

1. Defines correctly 14 words out of a list of 45 arranged in increasing difficulty.
2. Detects verbal absurdities such as the one that asserts that in an old graveyard in Spain there was found a skull believed to be that of Christopher Columbus when he was 10 years old.
3. Explains what has happened in a picture in which a messenger boy who has broken his bicycle is hailing a passing motorist.
4. Repeats five digits reversed.
5. Defines abstract words such as *constant* and *charity*.
6. Completes sentences with words omitted.

When these sets of items are accurately located at the correct year they may be used as points of reference. Thus a child who answers the items of year VI is solving 6-year-old problems. A child's mental capacity may, then, be derived directly from the scores on the test. The number of years and months he scores may be thought of as his *mental age*.

Mental Age

By means of mental age, it is possible to compare a child of any chronological age with the mental performance of the average child. As a consequence, it is possible to say about a child of 9 years of age that he has a mental age (M.A.) of 6. In 9 years of living his mental development has reached only that of an average 6-year-old child. He is retarded 3 years in his mental development. Let us consider the records of two children tested in the Terman-Merrill Revision.

The first child has a chronological age of 14 years and 11 months (usually written 14-11). Here is his record:

| Years | Months |
|-----------------|--------|
| VII (basal age) | 84 |
| VIII | 6 |
| IX | 2 |
| X | 4 |
| XI | 2 |
| Total | 98 |

M.A. = 8 years and 2 months (8-2)

$$\text{I.Q.} = \left(\frac{8-2}{14-11} \right) 100 = 57$$

Basal age is defined as that age on the test where all items are passed. Testing is frequently begun at a year under a child's chronological age.

The tester usually then proceeds down the scale until *all items at one age* are passed and up the scale until all items are missed.

The second child has a chronological age of 8-6. His test record follows:

| Years | Months |
|------------------|--------|
| VIII (basal age) | 96 |
| IX | 8 |
| X | 8 |
| XI | 6 |
| Total | 118 |

$$\text{M.A.} = 9-10$$

$$\text{I.Q.} = \left(\frac{9-10}{8-6} \right) 100 = 116$$

Intelligence Quotient (I.Q.)

The intelligence quotient, ordinarily called the I.Q., expresses the ratio between chronological age and mental age. As has been indicated in the two cases just described, it may be written

$$\text{I.Q.} = \frac{\text{M.A.}}{\text{C.A.}} \times 100$$

In the first child this becomes

$$\left(\frac{8-2}{14-11} \right) 100 = 57$$

In the second it becomes

$$\left(\frac{9-10}{8-6} \right) 100 = 116$$

The intelligence quotient indicates *both* the intelligence which an individual possesses *and his rate of growth*. Let us consider the accompanying table, which is of aid in interpreting all I.Q.s. It is derived from the studies of the Terman-Merrill Revision and is recommended by Dr. Merrill.¹

| I.Q. | |
|---------|----------------------|
| 140-169 | Very superior |
| 120-139 | Superior |
| 110-119 | High average |
| 90-109 | Normal or average |
| 80-89 | Low average |
| 70-79 | Borderline defective |

¹ Merrill, Maud A., "I.Q.s on the Revised Stanford-Binet Scale," *Journal of Educational Psychology* (1938) 29:641-651.

From this table, the I.Q. of 57 places our first child in the category of *mentally defective* while the second child's I.Q. of 116 places him in that of *high average*.

The I.Q. also indicates something about the *rate of growth*. Let us, for illustration, consider three I.Q.s: 50, 100, and 150. The rate of growth of the child of 50 I.Q. is about half that of the normal child. It takes such a child 2 chronological years to grow 1 mental year. At 8 years of age he has grown only 4 mental years. The child with the I.Q. of 100 grows 1 mental year during 1 chronological year. When he is 8 years old his mental age is also 8. The third child, with the I.Q. of 150, is growing at an accelerated rate. By the time he is 4 his mental age is 6 and when he arrives at 8 his mental age is 12. Moreover, these three children will continue to grow at *somewhat the same rate as they have grown*. *The I.Q. then gives us some indication of the rate of growth to be expected.*

Four characteristics of intelligence tests need to be kept in mind when attempting to understand them:

1. I.Q.s are *not inherited*. They are, as is every other aspect of mental life, the results of the interaction of inheritance and environment. Newman has shown that two identical twins who differed 13 years in education differed 24 I.Q. points.¹ Each of the pair had the same *genetic constitution*, but in one case there was not enough environmental stimulation to develop this capacity. A child from a poverty-stricken family who earns an I.Q. of 90 has more native capacity than a child with the same I.Q. from an excellent environment. Children deaf from birth frequently have low I.Q.s simply because they have been shut off from environmental stimulation. Let us once and for all abandon the idea that the I.Q. is inherited like the color of our eyes or the freckles on our skin.

2. I.Q.s are *not constant* but vary considerably within limits. Variation of I.Q.s may be due to the manner in which a test is given or scored, to the fact that they are derived from tests standardized on different populations, or even to the fact that one child cheated. An I.Q. of 100 obtained from the correct administration of a test would vary 4 or 5 points on the second giving. There are 99 chances in 100 that an I.Q. of 100 would not vary more than 15 points in the administration of two forms of the same test. The variations just discussed are those arising out of the process of measurement. Radical changes in environment or emotional maladjustments may produce greater variations than those described. On the other hand, we do not expect a child with an I.Q. of 50 ever to be normal or one with an I.Q. of 130 ever to recede to 100.

¹ Newman, Horatio H., *Multiple Human Births*. New York: Doubleday & Company, Inc., 1940.

Intelligence quotients remain within certain definable limits from year to year, but the limits are broad.

3. An I.Q. is *more valuable the nearer in time* it has been computed. An I.Q. computed for a 3-year-old is of very little value at age 6. The testing of very small children is fraught with many difficulties, *e.g.*, negativism. After year 6 or 7, the I.Q. stays more nearly the same, *i.e.*, its variation is less. For a sixth-grade teacher to have to depend on an I.Q. secured in the third grade is unfortunate indeed. If it has been computed while the child was in the fifth grade it is valuable.

4. Intelligence tests, except for performance tests, measure *verbal intelligence*. This means that a poor reader in the fourth grade will be penalized by giving him a group intelligence test. Poor reading then is frequently the cause of low scores on group intelligence tests.

If these matters are kept in mind, intelligence test scores and I.Q.s are the most useful types of information which can be collected. They indicate the child's present learning capacity and help the teacher in knowing what procedures are best for his continuing development.

Evaluation of the Terman-Merrill Revision

What of this latest test constructed after the Binet style— does it stand up above other tests? It does. Most workers believe it the best test of its kind ever constructed. It will undoubtedly be used more than any other individual tests, and yet there are those who believe that further improvement will come from other directions. They say, for example, "the new Stanford Revision is probably the last of the mental age scales" because its standardization is "laborious, rigid, and final."¹ Another criticism in the same direction is voiced by clinical psychologists who deal with individual cases frequently nervous in disposition. One worker² thinks the scoring by points is less cumbersome, that the form of the Terman-Merrill Revision is inconvenient and that the 45-word vocabulary test is dreadfully inadequate. She believes, furthermore, that to ask subjects to define words orally is an imposition in that the best subjects will not answer because they are satisfied only with dictionary definitions and hence keep silent. Then, too, the procedure whereby the subject is carried back until he is correct in all at that age level and forward until he misses all is a bad feature. It is bad because the test usually ends with a half dozen failures in succession and to some nervous subjects this series of failures is sheer torture.

¹ Freeman, F. N., *Mental Tests*, rev. ed. p. 106. Boston: Houghton Mifflin Company, 1939.

² Kent, Grace H., *The Nineteen Forty Mental Measurements Yearbook* (Oscar K. Buros, ed.), Item 1420. Highland Park N.J.: The Mental Measurements Yearbook, 1941.

On the other hand it is the opinion of one large clinic that has used this test on more than a thousand cases¹ that the new test is superior statistically in every way to the old test (1916 edition). It eliminates many objections of the old, it tests the brighter more effectively as they grow older. But it also has its weaknesses. The newer test takes 25 to 30 per cent more time to give than the Stanford-Binet. There is still too much emphasis on verbal material, especially in years VIII and XI. Many tests are misplaced for New York children, for example, and there is need in the case of clinical work for more flexibility of administration. Finally, the critics mention a weakness in basal age. The *basal age* is the age at which all the tests are passed. In scoring there are added to the score of the basal age additional mental months scored in the ages above the basal age. In the new test there may frequently be two basal ages or even at times three. For example, a child who is tested at the age of 10 passes all the tests at year X, misses one at year XI, and gets all tests right at year XII, thus both year X and year XII are the basal ages and the M.A. will be different depending on which one is used. Thus one investigator found, when 67 freshmen and 86 senior medical students were tested with the new revision, that the average number of basal ages for the freshmen was 1.5 and for the medical students, 2, 45 per cent of the freshmen had more than one base, and 56 per cent of the medical students likewise. The reason for this multiplicity of basal ages is that the mental-age growth from one year to the next is a very small amount indeed at years XIV, XV, etc. Year XIII seemed to be more difficult for these college students than either year XII or year XIV.

The criticism concerning the large number of verbal tests was met squarely by Terman and Merrill. They definitely tried to secure other tests which would stand up to their criteria for selecting tests. But only in rare cases were they able to discover useful performance tests. They believe that language enters inextricably into the upper levels of intelligence and to be able to think abstractly demands, in most cases, words. These authors undoubtedly would reply to the criticism concerning the small sample of words in the vocabulary test, that *it works* and, furthermore, that this test is not intended to test an individual's vocabulary but through his vocabulary to get an indication of his level of intellectual development. The multiple mental ages illustrate how difficult it is to get tests which depend on environments common to all. Change your environment sufficiently and the placing of your tests is immediately affected, so that a level of year XII is changed to year XIII, etc.

In conclusion, there is a *fundamental weakness in the Terman-Merrill Revision in Testing adults*. The difficulty arises in connection with the

¹ Krugman, M., "Some Impressions of the Revised Stanford-Binet Scale," *The Nineteen Forty Mental Measurements Yearbook*, op. cit., Item 1420.

concept of mental age after the year 15 or 16 is passed. It is a well-known fact that the differences between mental years decreases after age 12 or 13. Mental growth, in brief, slows down. The answer to the question as to when it ceases entirely has varied from 14 to 25 years. The experience gained in testing in the First World War indicated that the average age of mental maturity is 14. Terman in the original Stanford-Binet used 16. In the Terman-Merrill Revision the age of 15 is used. This means that if we used 16, the denominator of the intelligence quotient for any subject 16 years, 17 years, or 25 years old would always be 16. The concept of mental age, then, has only hypothetical meaning after age 15 or 16. (Wechsler,¹ for example, substitutes for the I.Q. an efficiency quotient. This author claims that an intelligence quotient must always refer a subject's score to the mean of *his* age.) This fact limits the effectiveness of the Terman-Merrill Revision for measuring intelligence after the age of 15 or 16. A second weakness consists of large variations in the standard deviations at various chronological ages.

The standard deviations on Form L, for example, vary from 12.5 at year 6, to 20 at year 12, and 20.6 at year 21½. This means that a child's I.Q. of 112.5 at year 6 would correspond to 120 at year 12. Terman and Merrill average these yearly differences and use a standard deviation of 16 at all ages. Variations in I.Q.s from year to year would thus be affected by the very manner in which the tests are constructed.²

POINT SCALES

The Wechsler-Bellevue Intelligence Scale

In 1939 there was published for the first time the Wechsler-Bellevue Intelligence Scale. This scale resembles in scoring the point scale of Yerkes *et al.*

This individual scale is suitable for subjects *who are 10 years of age and up*.³ It is particularly well suited for testing adults. This scale offers a serious challenge to all other tests of adult intelligence. It claims to measure the major part of what is contained in this definition: "*Intelligence is the aggregate or global capacity of the individual to act purposefully, to think rationally and to deal effectively with his environment.*"⁴

Its general form resembles very closely that of the group tests of intelligence. There are 10 test forms and one substitute test, a test of

¹ Wechsler, David, *Measurement of Adult Intelligence*, 3d ed., p. 46. Baltimore: The Williams & Wilkins Company, 1944.

² Terman and Merrill, *op. cit.*, p. 40.

³ A new test for children has now been constructed which extends the testing range to five years.

⁴ Wechsler, David, *Measurement of Adult Intelligence*, p. 3. Baltimore: The Williams & Wilkins Company, 1944.

vocabulary. These test forms or subtests are divided into two parts. Part I, which is verbal in nature, consists of five tests: (1) information, (2) comprehension, (3) digit span, (4) arithmetic reasoning, and (5) similarities, and an alternate test, vocabulary. Part II, which is a performance test, also consists of five parts: (1) picture arrangement, (2) picture completion, (3) block design, (4) object assembly, and (5) digit symbol.

Each of these subtests has a list of items of increasing difficulty from three form boards in object assembly to 25 items of the information test and 42 words to be defined. These subtests were chosen because they had proved their worth in the general statistical appraisal or else had been highly considered in clinical practice. They were kept in the scale because each of them correlated well with the test as a whole and because each contained items which an individual could have acquired from ordinary experience. Some idea of the value of each subtest may be had from Table 10.

TABLE 10. CORRELATION OF EACH SUBTEST WITH THE SCALE AS A WHOLE EXCLUSIVE OF THE TEST IN QUESTION
(Ages 20-34. $N = 355$)

| 1. Verbal | | 2. Performance | |
|--------------------|-----|--------------------------|-----|
| Information..... | .67 | Picture arrangement..... | .51 |
| Comprehension..... | .66 | Picture completion..... | .61 |
| Digit span..... | .51 | Block design..... | .71 |
| Arithmetic..... | .63 | Object assembly..... | .41 |
| Similarities..... | .73 | Digit symbol..... | .67 |
| (Vocabulary)..... | .85 | | |

The subtest "Similarities" shows the closest relation with the combination of the other tests, and that on object assembly the least. Interesting is the high relationship of .85 between vocabulary and the test as a whole. Terman valued vocabulary very highly; Wechsler is forced to do so.

In spite of Wechsler's criticism of the use of the time factor in the Terman-Merrill Test, the scores of five tests are dependent upon time. These are arithmetic reasoning, picture arrangement, block design, object assembly, and digit symbol.

Each subtest is scored and the sum of the correct items is brought forward and recorded in a table on page 1 of the test blank (Table 11). These scores are then transmuted to weighted scores and added up. These weighted scores are summed up under (1) verbal score, (2) performance score, and (3) total score. Tables are furnished by which an I.Q. may be attained for each of the three divisions.

The validity and reliability of this test are reported in the book, *Adult Intelligence*. The validity of the test is established first of all by

TABLE 11. SCORE SHEET OF WECHSLER-BELLEVUE TEST FOR AN INDIVIDUAL AGED 30

| TABLE OF WEIGHTED SCORES† | | | | | | | | | | SUMMARY | | |
|---------------------------|-------------|---------------|------------|------------|--------------|------------|---------------------|--------------------|--------------|--|--------------|---------------------------|
| Equivalent Weighted Score | RAW SCORE | | | | | | | | | TEST | R.S. | W.T.S. |
| | Information | Comprehension | Digit Span | Arithmetic | Similarities | Vocabulary | Picture Arrangement | Picture Completion | Block Design | Object Assembly | Digit Symbol | Equivalent Weighted Score |
| 18 | 25 | 20 | 14 | 23-24 | 41-42 | 20+ | | | 38+ | | | 18 |
| 17 | 24 | 19 | 17 | 21-22 | 39-40 | 20 | | | 38 | 26 | | 17 |
| 16 | 23 | 18 | 16 | 20 | 37-38 | 19 | | | 35-37 | 25 | 66-67 | 16 |
| 15 | 21-22 | 17 | 15 | 19 | 35-36 | 18 | | | 33-34 | 24 | 62-65 | 15 |
| 14 | 20 | 16 | 15 | 17-18 | 32-34 | 16-17 | | | 30-32 | 23 | 57-61 | 14 |
| 13 | 18-19 | 15 | 14 | 16 | 29-31 | 15 | | | 28-29 | 22 | 53-56 | 13 |
| 12 | 17 | 14 | | 15 | 27-28 | 14 | | | 25-27 | 20-21 | 49-52 | 12 |
| 11 | 15-16 | 12-13 | 13 | 13-14 | 25-26 | 12-13 | | | 23-24 | 19 | 45-48 | 11 |
| 10 | 13-14 | 11 | 12 | 12 | 22-24 | 11 | | | 20-22 | 18 | 41-44 | 10 |
| 9 | 12 | 10 | 11 | 11 | 20-21 | 10 | | | 18-19 | 17 | 37-40 | 9 |
| 8 | 10-11 | 9 | | 9-10 | 17-19 | 9 | | | 16-17 | 16 | 33-36 | 8 |
| 7 | 9 | 8 | 10 | 8 | 15-16 | 7-8 | | | 13-15 | 14-15 | 29-32 | 7 |
| 6 | 7-8 | 7 | 9 | 7 | 12-14 | 6 | | | 11-12 | 13 | 24-28 | 6 |
| 5 | 6 | 5-6 | | 5-6 | 10-11 | 5 | | | 8-10 | 12 | 20-23 | 5 |
| 4 | 4-5 | 4 | 8 | 4 | 7-9 | 4 | | | 6-7 | 10-11 | 16-19 | 4 |
| 3 | 2-3 | 3 | 7 | 3 | 5-6 | 2-3 | | | 3-5 | 9 | 12-15 | 3 |
| 2 | 1 | 2 | 6 | | 3-4 | 1 | | | 1-2 | 8 | 8-11 | 2 |
| 1 | 0 | 1 | | 2 | 0 | 1-2 | 0 | | 0 | 7 | 4-7 | 1 |
| 0 | | 0 | 5 | 1 | 0 | 0 | | | | 5-6 | 0-3 | 0 |
| | | | | | | | | | | VERBAL SCORE* | | |
| | | | | | | | | | | (VOCABULARY) | (| (74) |
| | | | | | | | | | | P. ARRANGEMENT | 15 | 13 |
| | | | | | | | | | | P. COMPLETION | 15 | 15 |
| | | | | | | | | | | BLOCK DESIGN | 34 | 15 |
| | | | | | | | | | | OBJECT ASSEMBLY | 23 | 14 |
| | | | | | | | | | | DIGIT SYMBOL | 56 | 13 |
| | | | | | | | | | | PERFORMANCE SCORE* | | 70 |
| | | | | | | | | | | TOTAL SCORE | | 144 |
| | | | | | | | | | | *Proration is necessary if four or six Verbal tests are given or four Performance tests. | | |
| | | | | | | | | | | VERBAL SCALE | 74 | 1.0. 133 |
| | | | | | | | | | | PERFORM. SCALE | 70 | 1.0. 131 |
| | | | | | | | | | | FULL SCALE | 144 | 1.0. 132 |

†Clinicians who wish to draw a "psychograph" on the above table may do so by connecting the appropriate raw scores; however, one must recognize the relative unreliability of these subtest scores when they are thus treated. (By permission of Williams and Wilkins Company, Baltimore)

reference to clinical practice. The author emphasizes the importance of agreement between the test score of an individual and his adaptation to environment as the most important evidence of the test's validity. Unless there is agreement in this instance the test fails.

It is this appeal to actual success in the clinic on which Wechsler places the greatest trust. He furnishes evidence to show that in case after case the Wechsler Bellevue I.Q. agreed more closely with the subject's life success than did I.Q.s from other tests. Furthermore, when correlations were computed between psychiatrists' recommendations of "commitment" or "noncommitment" to a state institution and I.Q.s achieved, the results were as follows:

| | |
|--|---------|
| Stanford Binet I.Q.s and psychiatrists' recommendations.. | r 33 |
| Wechsler Bellevue I.Q.s and psychiatrists' recommendations | 79 |

The evidence as submitted is greatly in favor of the Wechsler-Bellevue. An even better comparison between the two tests is obtained when their forecasting efficiency is compared. The predictive efficiency of a correlation of .33 is 5.6 per cent, that of a correlation of .79, 38.7 per cent.

While Wechsler does not believe that correlation with teachers' estimates of intelligence adds much to a test's validity, he did make comparisons between his I.Q.s and these estimates. The subjects were from the high school level. The coefficients were .43 and .52. In a similar comparison between the Stanford-Binet and teachers' estimates of intelligence the corresponding coefficient was .48. It is thus seen that no significant difference between the two tests appears in this case. Finally correlations are also furnished with older measuring instruments. With the Terman Merrill Revision, coefficients of .91, .62, .93, and .89 have been computed. With group tests, the coefficients are somewhat lower: Henmon Nelson, .81, Army Alpha, .74, A.C.E. (American Council on Education), .53; and Thorndike's C.A.V.D., .69 and .39.¹ It is clear that the Wechsler-Bellevue measures much the same sort of thing as does the Terman-Merrill Revision and correlates with group tests about like other individual tests.

The report on the reliability is not all that could be desired. The reliability is computed from the repetition of the same test at intervals of 1 month to 1 year. It is true that the reliability coefficient of .94 for both children and adults is adequate, but the number of cases used is definitely inadequate. Only 32 children between the ages of 10 and 13 were used in computing the coefficient, and 20 adults. Moreover, the correlation is computed by means of the rho formula, which is more unreliable than the standard Pearson product-moment method.

¹ *Manual*, p. 134.

The determination of norms was carefully done. The population used in this procedure consisted of 670 children between the ages of 7 and 16, from 50 to 100 at each age, and 1,081 adults between the ages of 17 and 70. There were from 50 to 195 adults at each age group, with a hundred or more at each group from 17 to 40 and fewer than 100 after 40. The securing of samples truly representative of the total population was attempted. Noting that in general there is a significant correlation between the Wechsler-Bellevue and the level of educational advancement achieved, comparisons were made with these levels as achieved by the population of the United States. There is some tendency for the standardizing population to be better educated than the average. For example, 5.10 per cent of the Wechsler-Bellevue group were college graduates, while the average for the nation is 2.93, the corresponding figures for illiterates were 2.55 and 4.69. In the Wechsler-Bellevue population 19.68 per cent are high school graduates or above, while in the population at large this percentage is 13.86. Moreover, 10 per cent more of the Wechsler-Bellevue group are elementary school graduates only, while 13 per cent more of the general population did some elementary school work but did not graduate. The population sample was composed of whites only, and therefore the test is not recommended for use in measuring subjects of other races.

Distinctive Features of the Wechsler Bellevue Scale

1. The Wechsler-Bellevue scale abolishes the use of the mental age but keeps the I.Q. It is held (a) that the M.A. is only a score, and (b) that its range is limited beyond a certain age (usually 15 or 16). The nature of the I.Q. is changed somewhat. In this test,

$$\text{I.Q.} = \frac{\text{attained or actual score}}{\text{expected mean score for age}}$$

It is thus a ratio between an individual's achieved score and the *mean* of the age group to which the individual belongs. It gives an individual's relative position in his own age group. For these reasons the I.Q. keeps the same meaning throughout life.

2. It is a point scale whose scores are transmitted into standard score units. This is not so distinctive as might at first appear. The Terman-Merrill Revision calculated the standard deviation of 16 to be used with Form L. An I.Q. of 116 in this latter test is 1 standard deviation above the mean.

3. It makes allowance for the gradual deterioration of intelligence with age. An illustration of this occurs when a score of 70 is considered.

A score of 70 on the full scale gives the following I.Q.s according to the age:

| Age | I.Q. |
|-------|------|
| 20-24 | 80 |
| 25-29 | 83 |
| 30-34 | 86 |
| 35-39 | 89 |
| 40-44 | 91 |
| 45-49 | 93 |
| 50-54 | 95 |
| 55-59 | 97 |

This is its most distinctive feature and by far its most important one. It constitutes a definite improvement over other individual scales.

4. The use of subtests whose scores are transmuted into standard scores makes it possible to know immediately in which area of intelligence the individual is weak or strong and to construct a profile if one is desired.

5. It allows for the computation of the I.Q. based either on verbal tests, on performance tests, and on both together. For poorly educated adults the I.Q. based on performance tests is of very great value.

The evidence as a whole clearly indicates that the Wechsler-Bellevue is the best instrument available for testing adult intelligence.

PERFORMANCE TESTS

Tests which lean rather heavily on the definitions of words and upon other verbal problems are decidedly unfair to those whose language development has been retarded for some reason or other. Deaf children immediately come to mind, as well as those who have been reared in socially isolated pockets or those whose education is much below par. Now and then, too, a child's environment has been so bookish and so verbal that he scores higher on a verbal test than he really has attained. As Spearman would say, "his *s* has become as important as his *g*."

Two points of view are extant concerning performance tests. In one of these, the performance test is coordinate and equal to the verbal test. On the one hand we have a sentence with words omitted, on the other, appear pictures with certain parts omitted. The other point of view regards the performance test as distinctly supplementary to the verbal test and as supplying a phase of planning and problem solving not encountered in the first instance. These performance tests ask you to do something about the problem. In a large picture you are, for example, to observe what book a boy has lost on the way to school and to select from the several pictures available that book with precisely the right color. This involves, of course, the understanding of the total import of

the picture, a keen observation of what was present in a previous picture, a recognition of what is not now there, and finally the selection of the correct picture. In many of the performance tests there is need of keen observation, and then of analysis and selection. In a simple form board the subject must perceive the size and shape of the opening and then select out of many that block which fits the opening exactly. Again, he must actually thread his way through a pencil maze whose imaginary walls cannot be crossed, but along whose imaginary road the subject must move his pencil to an imagined goal. Do such procedures involve the same sort of intelligence that is present in answering verbal problems? It is impossible to tell by introspective analysis. The only real way to solve our problem is through the aid of the coefficient of correlation.

Even when we use this coefficient we cannot be certain of the answer. The reason for this is that many correlations between performance tests and verbal tests have not reckoned on the C. A., which is related to both. If we compute the coefficient of correlation between the Pintner-Paterson series and Stanford-Binet we get a correlation in the neighborhood of .80. Based on this figure we could say that 64 per cent of the variance in the verbal test was associated with the variability of the performance test.¹ But we have failed to consider the fact that both tests are correlated highly with age. The r between C.A. and Stanford-Binet M.A. is close to .90; and between Pintner-Paterson and C.A. about .75. When the factor of age is "partialed out" (made constant) the true correlation between these two tests is reduced to .43 and their percentage of dependent variance is reduced to 18. If this line of argument is correct, then those students who believe the performance tests are supplementary to the verbal ones are correct. Another bit of evidence fits into this pattern. When the new Terman-Merrill Revision was being constructed the authors were very anxious to do away with that continuing criticism that the Stanford-Binet depended entirely too much upon language facility. They tried out several performance tests with that consciously in mind, but to no avail. These authors could find few performance tests which at the middle and upper levels of intelligence satisfied the criteria laid down for the construction of the test as a whole.

The Pintner-Paterson Scale of Performance Tests

These tests differ sharply from the Binet tests (1) in requiring actual manipulation of material to solve the problem, and (2) in not leaning

¹ This r^2 gives the percentage of the variance of the dependent variable which is associated with the independent variable. This interpretation is "a more general result than is the interpretation of r^2 as giving the percentage of elements in one test which are also in the other test." Garrett, Henry E., *Statistics in Psychology and Education*, 2d ed., p. 355. New York: Longmans, Green & Co., Inc., 1938.

too heavily upon the relations between words. Most of these tests demand observation, memory, and manipulation for their solution. The Pintner-Paterson scales consist of 15 different tests that are given separately and scored separately. Seven of these tests consist of some type of form board: Séguin, two-figure, five-figure, Casuist, triangle test, diagonal, and Healey Puzzle A (Fig. 29). These tests demand of the subject keen observation of the size and shape of holes and of cutouts that fit into those holes, and of the proper manipulation of the cutouts into the proper holes. In some tests there is needed a perception of the

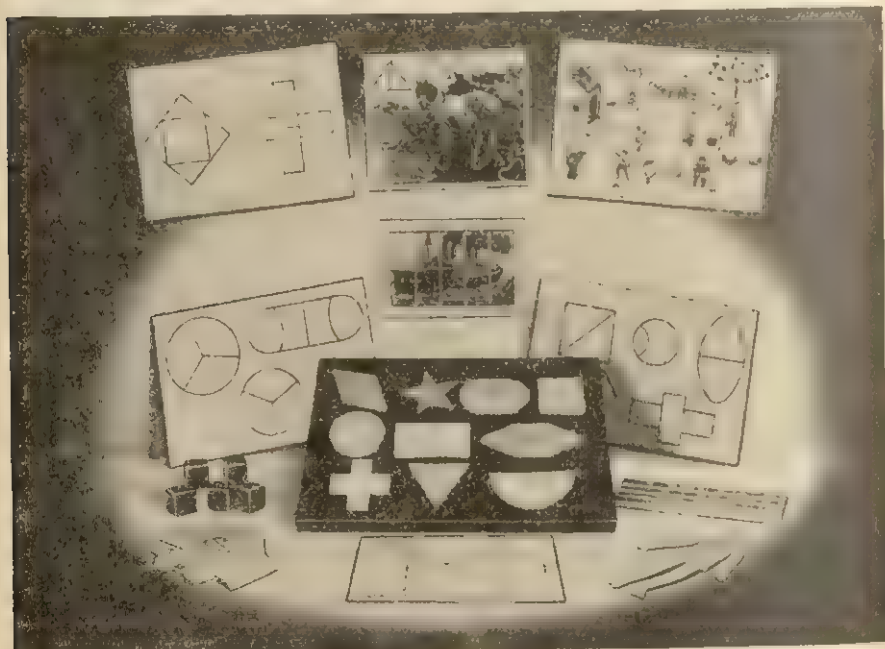


FIG. 29. Pintner-Paterson Performance Test, short scale. (By permission of C. H. Stoelting Company, Chicago.)

relation of each part of the materials to the whole problem. There are three tests—Mare and the Foal, Ship Test, and Healey Picture Completion I—which depend more upon understanding the problem as a whole and less upon the manipulation of the parts. In the ship test, for example, slices are made all the way through the picture of a ship. In the test, these parts of the picture are placed in an irregular order. If the parts are placed together correctly, a complete picture of a ship is the result. The other tests are difficult to classify, although the manikin and feature profile depend upon the same grasping of the total situation as do the completion tests. The substitution test is simply a technique for measuring the speed of learning—*i.e.*, the placing in each form of a

number which had already been decided upon and arranged in a key at the top of the test. The cube imitation test consists of five 1-inch cubes. Four of these are placed on a table and the fifth one is used to tap the tops of the others in a definite pattern. The subject watches closely, takes the cube, and taps out the same pattern that the experimenter has just demonstrated. The patterns then become more complex. The final test, the adaptation board, consists of a board with four round holes. Three of these are 6.8 centimeters in diameter and the fourth, 7 centimeters. One block fits exactly into the large hole (a fact demonstrated to the child). The whole board is then placed in four different positions. The child puts the block each time into the hole that it fits exactly.

For more convenient testing these 15 tests have been reduced to 10 by omitting the triangle test, diagonal test, Healey Puzzle A, substitution test, and adaptation board.¹ By reducing the size of some of the form boards the whole test may be conveniently carried in a small case. The short scale is probably to be preferred to the long one, both from convenience and because of fewer form boards.

The age range of this test is from 4 to 15 years. There are no tests which are discriminative over this total range. Two or three form boards are of little value after the M.A. of 10, and the feature profile has no value until after 10.

Arthur's Point Scale of Performance Tests

This arrangement of tests is composed of two forms. Form I is made up of eight tests from the just-mentioned Pintner-Paterson series, together with the Kohs Block Design Test and the Porteus Mazes. Form II is made up of Healey Picture Completion II, the Porteus Mazes, and the Kohs Block Design Test, along with five tests selected from the Pintner-Paterson series. These tests were restandardized on the basis of records secured from 1,100 school children, ages 6 to 16.²

Goodenough "Drawing a Man" Scale

Here is another performance test which requires no apparatus and usually not more than 10 minutes of the subject's time. Its instructions are straightforward and simple: "Make a picture of a man. Make the very best picture that you can." The score bears no relation to artistic

¹ Hildreth, Gertrude, and Rudolph Pintner, *Manual of Directions for Pintner-Paterson Performance Tests, Short Scale*. Bureau of Publications, Teachers College, Columbia University, New York, 1937.

² Arthur, Grace, *A Point Scale of Performance Tests*. New York: Commonwealth Fund, Division of Publication.

ability but only to the number of parts which the subject enters. Legs, arms, eyes, fingers, nose, mouth, etc. each one counts a point, 51 points in all. The test covers the range from 3 to 13 years and works best between the ages of 4 and 10. Tables are furnished so that these points scored may be transmuted into mental age in the usual way. The medians on these tables were computed from the scores of nearly 4,000 children. The reliability of this test was .94 when the data were based on a retest of 194 first-grade children. For ages 5 to 10 taken separately, the reliability coefficient was .77 on the average. Girls do better than boys on this test, though the sex differences are not marked. The test was standardized upon children who were at age in the grade tested. A man was chosen to be drawn because his clothing was more uniform than that of a woman or girl. Those points were selected for scoring which showed (1) a regular and rapid increase in percentage of children succeeding at successive ages, (2) a clear difference between performances of children at the same age but in different school grades. The points to be given are carefully described and illustrated.

THE MEANING OF INTELLIGENCE

There has not been up to the present time any general agreement concerning the meaning of intelligence. It seems to the author that the essence of intelligence is contained in one aspect of Binet's definition. Binet defined intelligence as (1) the ability to take and maintain a given mental set, (2) the capacity to make adaptations for the purpose of attaining the desired end, and (3) the power of self-criticism. *The capacity to make adaptations for the purpose of attaining the desired end* is at the very heart of the meaning of intelligence and the author believes it is very nearly the meaning espoused by many psychologists.

Some years ago (1921) a number of psychologists were asked to express their individual opinions as to what each thought intelligence was.¹ There were over 20 replies. Let us compare the definition of "adaptation" with a few of these definitions. Colvin's definition as "the capacity to learn" is simply another way of saying adaptation for the purpose of attaining the desired end. Indeed the last part of the preceding sentence could be done away with, since hardly ever would adaptation occur unless it was directed toward a desired end.

Let us look backward a moment to the definition of intelligence written by Wilhelm Stern who, you remember, gave us the I.Q. "Intelligence is a general capacity of an individual consciously to adjust his thinking to new requirements." "It is a general mental *adaptability* to

¹ "Intelligence and Its Measurement" (symposium), *Journal of Educational Psychology* (1921) 12:123-147, 195-216.

new problems and conditions of life." A few other definitions much like this one will be given. Woodworth says, "He has to see the point of the problem now set him, and to *adapt* what he has learned to the novel situation." Wells's definition approaches very closely these others: "Intelligence means precisely the property of so recombining our behavior-patterns as to act better in novel situations." Of course there are degrees of adaptation. If an individual adapts well he has more intelligence than if he adapts poorly.

Another group of eminent psychologists places a slightly different emphasis upon what intelligence is, and yet the author believes all of them can be subsumed under one caption—*degrees of adaptation*. Thorndike, for example, defines intelligence as intellect, "as the power of good responses from the point of view of truth or fact." The emphasis here is upon the sagacity with which an individual adapts. He has more intellect in proportion as he selects the responses poorly or well. Ballard's definition is similar to the one given above: "The relative general efficiency of minds measured under similar conditions of knowledge, interest, and habituation." General efficiency for what? For making adequate adaptations to new situations. Not greatly different is Pintner's definition: "We must remember that intelligence is merely an evaluation of the efficiency of a reaction or group of reactions under specific circumstances." But what are the bases of evaluation if they are not the adaptation to a situation? If the situation is well adapted to, we give a high value to it, if not, a low one. Finally, let us look at F. N. Freeman's definition: "Psychologically, degrees of intelligence seem to depend on the facility with which the subject matter of experience can be organized into new patterns. This rearrangement of thought material is what characterizes particularly the higher mental processes." The organization of subject matter of experience into new patterns is most certainly adaptation at a higher level. An individual meets a problem which is complex and involved. He brings to bear his past experience, adjusts and arranges it, selects from it those facts which help him meet the present problem, and in this manner *adapts* to the problem. In proportion as he adapts well, he is intelligent. Finally, Terman's definition on first reading does not fit into the concept of adaptability. Terman defines intelligence as the "capacity for abstract thinking." This definition is probably meant to emphasize only the highest level of intelligence. As a matter of fact Terman says that simple motor activity at the pick-and-shovel level involves almost no intelligence. The representative level has a little more because, at this level, an individual can nurse you by carrying out the doctor's directions, or a builder can construct a house from the plans furnished him. The really intelligent man is none of these. He can think abstractly. He can plan you a house, in-

vent for you a preventive serum, or develop mathematical symbolism. He is the intelligent man. The reason that the ditchdigger or the nurse is intelligent, however, is because they can meet situations not before met with. If they can meet new situations in a way which will solve them satisfactorily they act intelligently. The scientist, too, differs from the ditchdigger and the nurse in that he adapts adequately to a complex situation. If one wants to restrict intelligence to the capacity to use that attribute of many situations which makes them alike, although different on the surface, and to use that attribute to interpret other situations, *i.e.*, to think abstractly, he is at liberty to do so. It seems a trifle restricted in conception to think of intelligence as that capacity to adapt by thinking abstractly. Surely this is the most successful of all adaptations and those who can think abstractly undoubtedly do possess a very high form of intelligence. It is submitted that intelligence in its very essence as used by competent psychologists in the great majority of cases is *adaptation to meet a desired end*.

It is clear that adaptation might refer to changes in the individual while the outside situation remained static. As a result there would be adaptation just as wheat or corn may be adapted to a very cold climate. On the other hand, the adaptation might be in the environment while the individual remains the same. Neither of these conditions usually exists. Generally speaking, there is a problem to be solved which arises out of a situation or a field of forces. To solve such a problem adaptation may be made of its component materials. However, the plan of change must have been evolved by the individual so that in a way he has adapted conditions around him to solve the problem. The successful solution of the problem would be evidence of his power of adaptation. Intelligence, then, varies in amount in proportion as the end is a complicated one or a simple one. A tiny child shows some intelligence when he adapts to a simple form board by placing the appropriate block in its hole. An older child is showing far more adaptation when he can figure out the probable height of a tree the next year after knowing that in four previous years the tree was 8, 12, 18, and 27 inches tall respectively. In life itself high intelligence is shown by an officer's successful handling of a problem in logistics which he has never met before. And how unintelligent such an one is regarded as being, if he keeps trying a memorized procedure which does not solve the present problem!

SUMMARY

The movement for the measurement and evaluation of intelligence arose both out of the scientific interest in individual differences and out of the practical problems of educating the backward and the feeble-minded. Its prime mover was Alfred Binet, who developed the first

standardized intelligence test. He also was the first to introduce the scientific meaning of mental age. The 1908 revision of the Binet-Simon tests was translated by Goddard, adapted to American children, and standardized upon a large number of them. Four revisions developed in the United States: (1) the Stanford Revision, (2) the Kuhlmann Revision, (3) A Point Scale for Measuring Mental Ability, and (4) the Herring Revision of the Binet Scales. Each of these scales has its advantages. The Terman-Merrill Revision of the Stanford-Binet is the most recent of these and probably the most satisfactory of all for testing children.

The Wechsler-Bellevue Intelligence Test, first published in 1939, resembles in general form a group test of intelligence in that each subtest contains similar items and its scores are in points but continues the use of the I.Q. However, the I.Q. of this test is the expression of a relation between an individual's score and the average score of his age group.

The recognition of the preponderant influence of language on the scores derived from the Binet tests led to the construction of performance tests. It was seen that the claims of these tests rested on the proposition that not all of intelligence is made up of verbal relations. Two views were introduced: (1) that performance tests were coordinate with the verbal tests, that they are another procedure to get at and measure the same mental traits; (2) that performance tests were subordinate or ancillary, adding a necessary and neglected part to the score furnished by the verbal tests. The Pintner-Paterson Scale of Performance Tests, Arthur's Point Scale of Performance Tests, and the Goodenough "Drawing a Man" Scale were described.

Along with the development with the instruments of measurement has appeared an interest in understanding more adequately the very nature of intelligence. Several definitions formulated by competent men have been introduced into the discussion. The notion of *adaptability* has been put forward as a definition which contains the elements of many others and perhaps the essential characteristic of intelligence.

QUESTIONS AND EXERCISES

1. Describe the two types of interests which led to the construction of the first tests of intelligence.

2. Explain the events which caused a setback to the early interest in test construction in the United States.

3. Who first introduced the Binet tests into the United States. What was this psychologist's major interest?

4. Criticize and evaluate the present author's concept of intelligence.

5. Summarize the leading changes introduced by Terman in the Stanford Revision; by Terman and Merrill in their 1937 revision.

6. Place on one side of a page the favorable facts concerning the Terman-Merrill Revision and on the other the unfavorable facts. Which seem to you to carry most weight?

7. Compare the leading characteristics of the Terman-Merrill Revision

with those of the Wechsler-Bellevue.

8. How does the Wechsler-Bellevue test provide for the gradual decrease of intelligence with age?

9. a. Distinguish between a performance test and a verbal test.

b. Describe the main features of the Pintner-Paterson Scale of Performance Tests.

10. Why has the Arthur Point Scale of Performance Tests been called the most useful of the performance tests?

BIBLIOGRAPHY

Books

ARTHUR, GRACE: *A Point Scale of the Performance Tests*, 2d ed. New York: Commonwealth Fund, Division of Publication, 1943.

FREEMAN, FRANK N.: *Mental Tests*, rev. ed. Boston: Houghton Mifflin Company, 1939.

GOODENOUGH, F. L., J. G. FOSTER, and M. J. VAN WAGENEN: *The Minnesota Pre-school Tests*. Minneapolis: Educational Test Bureau, 1932.

GOODENOUGH, FLORENCE L.: *The Measurement of Intelligence by Drawings*. Yonkers, N.Y.: World Book Company, 1926.

———: *Mental Testing: Its History, Principles, and Applications*. New York: Rinehart and Company, 1949.

HERRING, JOHN P.: *Herring Revision of the Binet-Simon Tests*, Examination Manual, Form A. Yonkers, N.Y.: World Book Company, 1931.

HILDRETH, GERTRUDE, and RUDOLPH PINTNER: *Manual of Directions for Pintner-Paterson Performance Tests, Short Scale, Ages 4 to 15*. New York: Bureau of Publications, Teachers College, Columbia University, 1937.

KENT, GRACE H.: *Nineteen Forty Mental Measurements Yearbook*, (Oscar K. Buros, ed.), Item 1420. Highland Park, N.J.: The Mental Measurements Yearbook, 1941.

KUHLMANN, F.: *Tests of Mental Development*. Minneapolis, Minn.: Educational Test Bureau, 1939.

PETERSON, JOSEPH: *Early Conceptions and Tests of Intelligence*. Yonkers, N.Y.: World Book Company, 1925.

PINTNER, RUDOLPH: *Intelligence Testing, Methods and Results*. New York: Henry Holt and Company, Inc., 1931.

PORTEUS, S. D.: *The Maze Test and Mental Differences*. Vineland, N.J., Smith Printing and Publishing House, 1933.

SPEARMAN, CARL: *The Abilities of Man*. New York: The Macmillan Company, 1927.

STUTSMAN, RACHEL: *Mental Measurement of Pre-school Children*. Yonkers, N.Y.: World Book Company, 1931.

TERMAN, LEWIS M., and MAUDE A. MERRILL: *Measuring Intelligence*. Boston, Houghton Mifflin Company, 1937.

THORNDIKE, EDWARD L.: *The Measurement of Intelligence*. New York: Bureau of Publications, Teachers College, Columbia University, 1926.

THURSTONE, L. L.: *Primary Mental Abilities*, Psychometrika Monograph, 1938.

WECHSLER, DAVID: *Measurement of Adult Intelligence*, 3d ed. Baltimore: The Williams & Wilkins Company, 1944.

WELLMAN, BETH L.: *The Intelligence of Pre-school Children as Measured by the Merrill-Palmer Scale of Performance Tests*, Studies in Child Welfare, Vol. 15, No. 3 (University of Iowa Studies, New Series, No. 361). University of Iowa, 1938.

YERKES, ROBERT M., and JOSEPHINE CURTIS FOSTER: *A Point Scale for Measuring Mental Ability*. Baltimore: Warwick and York Incorporated, 1923.

Articles

BERNREUTER, ROBERT G., and CHARLES H. GOODMAN: "A Study of the Thurstone Primary Mental Abilities Tests Applied to Freshmen Engineering Students," *Journal of Educational Psychology* (1941) 32:55-60.

FORREST, RUTH: *A Study of the Prognostic Value of the Merrill-Palmer Scale of Mental Tests and the Minnesota Preschool Scale*, unpublished master's thesis, University of Pittsburgh, 1939.

"Intelligence and Its Measurement" (symposium), *Journal of Educational Psychology* (1921) 12:123-147, 195-216.

MACMURRAY, DONALD: "A Comparison of the Intelligence of Gifted Children and of Dull-normal Children Measured by the Pintner-Paterson Scale, as against the Stanford-Binet Scale," *Journal of Psychology* (1937) 4:273-280.

MERRILL, MAUD A.: "The Significance of I.Q.'s of the Revised Stanford-

Binet Scales," *Journal of Educational Psychology* (1938) 29:641-651.

MITCHELL, MILDRED B.: "The Revised Stanford-Binet for University Students," *Journal of Educational Research* (1943) 36:507-511.

———: "Irregularities of University Students on the Revised Stanford-Binet," *Journal of Educational Psychology* (1941) 32:513-522.

SHARPE, S. E.: "Individual Psychology. A Study in Psychological Method," *American Journal Psychology* (1898-1899) 10:329-391.

WISSLER, CLARK: "The Correlation of Mental and Physical Tests," *Psychological Monographs* (1901) Vol. 3, No. 6.

CHAPTER 15

Group Tests of Intelligence

THE DEVELOPMENT OF GROUP TESTS

However successful competent workers were in constructing adequate intelligence tests, these instruments could not achieve their widest usefulness as long as it took the full time of a well-trained psychologist to administer the test to each individual. Only when large numbers of subjects could be tested at one sitting could the intelligence test reach its widest usefulness. It was the advent and development of the group intelligence test which brought about this condition.

Group tests of intelligence were slow in being developed because they were opposed by psychologists. Some authors have said that it took a great war to develop and popularize group tests of intelligence. It is undeniable that no group test of any consequence had advanced beyond the experimental stage before 1917. The reason for the slow development of group tests may now be explained.

It seemed to psychologists that there were definite advantages of the individual test. In the first place, the tester could adapt his test more certainly to the individual peculiarities of the subject such as negativism, scattering of attention, or lack of self-confidence. In the case of negativism, the skillful tester could get the child to solve a performance test before taking up the verbal problems. He could call the child back to his task by a variety of remarks and improve his lack of self-confidence by encouraging him after each test. "You are doing fine," "keep it up," and "you are doing well" are exhortations frequently used for encouragement. Then, too, the directions could be modified slightly or repeated until there was no question in the tester's mind concerning the child's understanding of the problem. In this manner, a child could be constantly motivated so that he did not attack one problem with cheerfulness and alacrity and another with gloom. One of the most difficult problems of the expert individual tester is this problem of rapport with the subject. Then too, there are cases of emotional maladjustment when the child simply refuses to take the test, in which case there is nothing to do but to try another time. Finally, many psychologists found in the testing situation an unusual opportunity for observing the emotional

reactions and work habits of the subject. Ratings were made of the willingness of the subject to cooperate, his self-confidence, his social consciousness, and his ability to keep his attention on his work. Even check lists have been provided for the purpose of collecting additional information about the personality adjustment to the testing situation. Certainly all these facts enter into the interpretation of whatever score is received.

How, then, can the group test compete at all with the individual technique of testing? The group test weathered the storm of criticism because it worked. After the age of 6 or 7 the limitations of group tests previously mentioned do not seem to affect the score a great deal.

Generally speaking, elementary school, high school, and college students are willing to take the test. Certainty of understanding is assured by stating the problem, illustrating it, and then having the student try a fore-exercise himself. The skillful tester watches closely for wandering of attention and when it occurs immediately steps quietly to the child and encourages him to work on the test or warns him that he has only a little time left. There is even a slight advantage residing in the group test when some self-conscious children are tested. Some of them become more self-conscious when a tester asks them oral questions. When, however, they are sitting in a class with others they lose themselves in the group and really make a better showing. However, some pupils refuse to cooperate and score very low or zero. Any child who scores very low or zero must be tested with another group test or with an individual test. One new difficulty presents itself in group testing—that of cheating. Clever testers have an answer for this problem. They stagger the tests so that no two children sitting side by side will be working on the same test; one of them will be at work on Form A and another on Form B.

The final clincher in this argument came when the reliability and validity of group tests were found to be satisfactory.

THE ARMY ALPHA AND ARMY BETA

The first group test grew immediately out of the exigencies of army need. In the First World War the army officers discovered that many draftees were mentally unfit for military service. They wanted an instrument that would sift out these men quickly without the long expensive procedure of trying them out in situations where they would fail. Some companies would be found ready to proceed to the front while others of the same regiment would be far behind in their efficiency. It was important to obtain well-balanced companies and regiments who were nearly at the same level in their mastery of army technique. It was just as important to select bright young men for officer material and for

other types of training. These were some of the uses to which a test could be put. But what of the test itself?

The committee of psychologists charged with the construction of this test wanted a test widely varying in difficulty - easy enough so that the stupidest could score something, and difficult enough to challenge the brightest ones. They needed, too, an instrument which was easily and accurately scored, would not take too long a time to administer, and would be interesting. To prevent cheating when a test was given, they thought that there should be several equivalent forms. With these thoughts in mind they discovered that Arthur S. Otis, then of Stanford University, had begun the construction of a group test of intelligence. This material was immediately made available to the army psychologists. Other types of tests were discovered and along with the Otis material assembled into what was known as Examination *a*. This test, consisting of 10 subtests, was finally revised, and reduced to eight tests, and labeled Army Alpha.

The Army Alpha Test was composed of eight subtests with a number of items under each:

- Test 1. Attention span
- Test 2. Arithmetic reasoning
- Test 3. Practical judgment
- Test 4. Same—opposite
- Test 5. Disarranged sentences
- Test 6. Number series
- Test 7. Verbal analogies
- Test 8. Information

As group tests of intelligence are investigated it will be clear that these subtests enter into their construction. In fact, extended studies of Army Alpha have demonstrated that it was and is a good test of intelligence.

The extensive use of the test in colleges for purposes of prediction of school success, for comparison of freshmen scores with army scores, and for assessing the various divisions of a university will be treated under the *uses* of intelligence tests.

Army Alpha required its subjects to be able to read easily and well in order for it to be a test of intelligence. Unless these conditions were realized failure resulted. During the First World War there were about 25 per cent of the draftees who for various reasons were functionally illiterate in the English language. It was necessary, therefore, to develop a test which made small demands upon the understanding of English. The test growing out of these demands was Army Beta. The tests consisted of tracing pathways through mazes, estimating the number of cubes in a drawn pile, completing a pattern of crosses and zeros arranged in a pattern, substituting symbols for numbers, recognizing sameness or

difference in a list of paired numbers, discovering the parts of pictures that were wrong, and with a pencil dividing up larger areas into which smaller areas would fit.

The score for the total test is a summation of the number of items correct in each test, with one exception. In Test 5 the score is one-third of the total right. As the test was used more and more it was discovered that instructions demonstrated on the blackboard and explained in gesture and pantomime were open to considerable variation in giving. The test did not prove as reliable or as valid as Alpha. There is a modern revision which uses only oral directions. The test was a forerunner of those tests of intelligence for little children which do not require a mastery of the written language. The Beta test, just as was the case with the performance tests, threw new light on the intelligence of those who have language handicaps.

THE PINTNER GENERAL ABILITY TESTS

A capital illustration of the group test is the series of tests called the Pintner General Ability Tests: Verbal Series. There are four different tests in this series: (1) the Pintner-Cunningham Primary Test, to be used from kindergarten through the first half of grade 2, (2) the Pintner-Durost Elementary Test, to be used from the last half of grade 2 through the first half of grade 4, (3) the Pintner Intermediate Test, suitable for last half of grade 4 and grades 5 through 8, and finally (4) Pintner Advanced, which is much like Pintner Intermediate Test but is more advanced and suitable for use with grade 9 through adult levels. The procedures used for the construction and standardization of this series of tests may be taken as examples of the available group tests of intelligence.

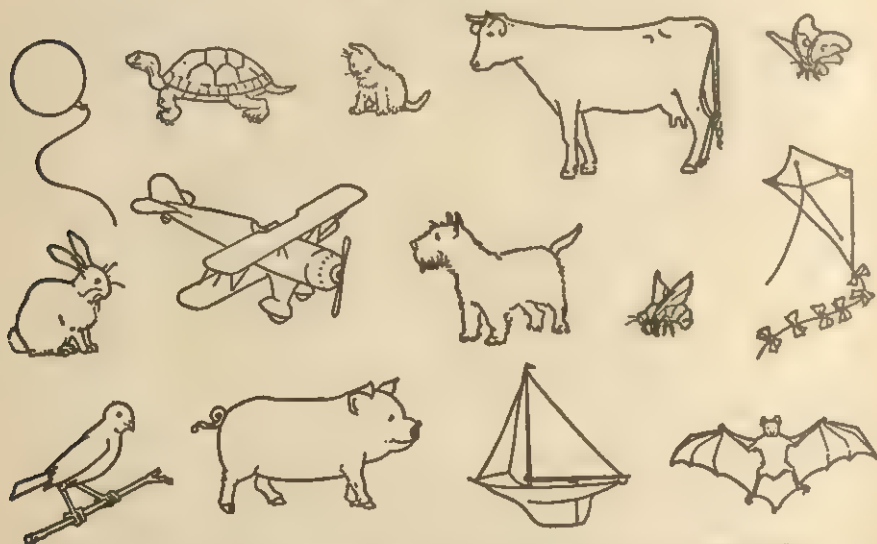
Figure 30 shows a page from the Pintner-Cunningham Primary Test, Form A.

Throughout this series great effort has been exercised to select only those items and subtests that had already proved their worth as efficient indicators of intelligence. In the tests for the little children who had not yet learned to read or who as yet read rather haltingly, dependence was placed on pictures. For example, in the Pintner-Cunningham Primary Test the children were asked to mark the pictures of objects that were alike in some way, or to mark what goes up in the air, or again to mark the prettiest of three pictured faces. Figure 30 shows one of the pictures and the instructions. The Pintner-Durost also uses pictures in Form I, Picture Content, by means of which ideas can be registered. In the intermediate and advanced tests eight subtests are used which experience had shown were probably the best of all. These are vocabulary,

logical selection, number sequence, best answer, classification, opposites, analogies, and arithmetic reasoning.

The validity of each test has been carefully studied and the results published for inspection. Correlations have been computed with the Stanford-Binet in the case of each of the four tests. In the case of every test this figure is about .80. In addition each test is correlated with many other evidences of intellectual progress. The Pintner-Cunningham test thus correlates well with measures of ability to succeed in school, and especially well with tests of reading. The intermediate and advanced

TEST 1



Total number right..... Total number wrong..... Score ($\frac{R-W}{9}$).....

FIG. 30. Pintner-Cunningham Primary Test. "Mark the things that go up in the air." (By permission of World Book Company.)

tests have coefficients of .70 or above with standard achievement tests or with other standardized tests of intelligence.

Probably in no test has the reliability been computed with greater care than is the case with this series. In all cases of reliability, coefficients computed from a range of one year have been furnished. It is a well-known fact that the restriction of the range reduces the magnitude of the coefficient. An illustration of this appears in the Pintner-Cunningham test, for in this test the reliability based on scores of pupils drawn from one grade varies from .83 to .89 but goes up to .94 when the pupils are drawn from members of the kindergarten, the first grade, and the second grade, a much wider range. Computations based on one grade or one age conform to the strictest canons of statistical accuracy. Relia-

bility correlations, based on pupils of one age or on one grade, are as follows:

| | |
|----------------------------|-----|
| Pintner-Cunningham..... | .89 |
| Pintner-Durost | |
| I. Picture Content..... | .85 |
| II. Reading Content..... | .95 |
| Pintner, Intermediate..... | .94 |
| Pintner, Advanced..... | .85 |

The standardization of these tests was adequate. The populations on whom the norms were established were studied for their representativeness and normality. One of the cities whose children's scores were used in the establishment of norms was shown to be an average American city. In the standardization of the intermediate test 100,000 tests of children representing both urban and rural populations were used.

There are several important features of this series of group intelligence tests. In the first place, reference has already been made to the use of standard scores. On page 364 it was shown that the Wechsler-Bellevue uses this type of score in computing the I.Q. In like manner standard scores are used with these tests to compute the I.Q.

TABLE 12. USE OF STANDARD SCORES IN COMPUTING I.Q.*
(Standard score norms corresponding to each age value. Intermediate and advanced tests, Forms A and B, regular edition.)

| Months | Years | | | | | | | | | | | | | |
|--------|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 0 | 101 | 113 | 124 | 134 | 143 | 150 | 158 | 164 | 171 | 177 | 182 | 187 | 191 | 195 |
| 1 | 102 | 114 | 125 | 135 | 143 | 150 | 158 | 165 | 171 | 177 | 182 | 187 | 192 | 196 |
| 2 | 103 | 115 | 126 | 136 | 144 | 151 | 159 | 165 | 172 | 178 | 183 | 188 | 192 | 196 |
| 3 | 104 | 116 | 127 | 136 | 145 | 152 | 160 | 166 | 172 | 178 | 183 | 188 | 192 | 196 |
| 4 | 105 | 117 | 127 | 137 | 145 | 152 | 160 | 166 | 173 | 179 | 184 | 188 | 193 | 197 |
| 5 | 106 | 118 | 128 | 138 | 146 | 153 | 161 | 167 | 173 | 179 | 184 | 189 | 193 | 197 |
| 6 | 107 | 119 | 129 | 138 | 146 | 154 | 161 | 167 | 174 | 180 | 184 | 189 | 193 | 197 |
| 7 | 108 | 120 | 130 | 139 | 147 | 155 | 162 | 168 | 174 | 180 | 185 | 190 | 194 | 198 |
| 8 | 109 | 121 | 131 | 140 | 148 | 155 | 162 | 169 | 175 | 180 | 185 | 190 | 194 | 198 |
| 9 | 110 | 122 | 131 | 141 | 148 | 156 | 163 | 169 | 175 | 181 | 186 | 190 | 194 | 198 |
| 10 | 111 | 123 | 132 | 142 | 149 | 157 | 163 | 170 | 176 | 181 | 186 | 191 | 195 | 199 |
| 11 | 112 | 123 | 133 | 142 | 149 | 157 | 164 | 170 | 176 | 181 | 186 | 191 | 195 | 199 |

* From Pintner's manual for administering and scoring the intermediate and advanced test.

Table 12 may be used to illustrate the use of the standard score, the M.A. and the I.Q. John, a boy 12 years and 6 months old, has been tested on the Pintner intermediate test and has earned a median standard score (since there are eight tests, each of which is reported in a standard score, the representative score would be their median) of 170. By looking at the table we can see that a child 12 years and 6 months would, if he were just normal, make a score of 154. Had he made a score of exactly 154, his I.Q. would then have been 100. If, as in this case, his score is more than the score norm for his age (here 154) the difference between what is normal and what the subject received may be added algebraically to 100. In this case, then, the I.Q. would be $100 + (\text{obtained score} - \text{norm for age})$ or $100 + (170 - 154) = 116$.¹ We can now compare this I.Q. with one computed in the usual way. John's chronological age (C.A.) is 12-6; his mental age (M.A.), secured by looking under median standard score for 170, is 14-10. His I.Q. computed in this manner therefore, is $14-10, 12-6$, or 119. You will note that this computation is 3 points larger when derived in the usual way. This is the exact procedure for years 11 and 12. For the other years there are slight modifications in scoring which are already worked out and made available in a table.

Other features of the Pintner Verbal Series are:

1. In the upper grades, all instructions are given before the subject starts to work. He works straight on through, unless he takes more than the allotted time for a single division, in which case the experimenter says, "Even if you have not finished test one, go on to test two," etc.
2. A profile may be drawn from the standard scores secured from each of the eight tests. This profile enables the experimenter to analyze the total score into eight divisions and to see immediately the areas of strength and weakness.

The Pintner Verbal Series in its selection of items, its manner of securing validity, its precision in calculation of reliability, its standardization based on a representative population, and in its use of the standard score is a worthy development from Army Alpha.

KUHLMANN-ANDERSON INTELLIGENCE TESTS

Another example of a test series is the Kuhlmann-Anderson Intelligence Tests. This well-known set of tests appeared first in 1927 and at the present (1951) has had five revisions. Altogether there are 39 tests

¹ It is interesting to compare this result with the procedure using the S.D. suggested by Terman and Merrill in *Measuring Intelligence*, p. 42 (Boston: Houghton Mifflin Company, 1937). The standard score there used is derived from an S.D. of 16, just the same as that used here for the median standard score. In the Terman-Merrill procedure a person whose I.Q. is 116 is just one S.D. above the average

which were selected from 100 after preliminary trials. These tests are arranged into nine batteries with ten tests in each battery. There are two first-grade batteries, one for the first semester of the first grade and the other for the second. Batteries are arranged for each of the grades from grade 2 through grade 6; one for grades 7 and 8; and finally one extending from grades 9 to 12. Each battery is made by including a few of the tests found more difficult at the preceding level and adding suitable new tests. In this manner the 39 tests are distributed into nine batteries.

The standardization of the tests has been carefully done. More than 30,000 Minnesota children, representative of the general population, were used to ascertain and check the median mental-age scores. Moreover, the original norms were based on at least 350 nonselected children at each age. One of the unique features of this series is that each test, made up of 6 to 24 items, is standardized separately. The M.A. then, is taken as the median M.A. of the 10 which the subject secures in each battery. This arrangement whereby each test is standardized separately has elements of strength. In the first place, a new test can be added with very little difficulty. One more M.A. may simply be put into the total pool and the median computed. Moreover, since each subject earns 10 different M.A.s one may compute an average or standard deviation from their medians. In this manner the variabilities of different subjects may be compared. Or again, the profile of the individual's scores received from the 10 tests may be used to discover whether the level of the test has been correctly chosen to correspond with the mental level of the subject. Though the procedure is not recommended by the authors, one might use this profile to secure an analyzed intelligence score. A subject might thus stand high in arithmetic reasoning and low in analogies, or high in copying visual forms and low in discovering the opposites to words.

Validity

The authors' procedure to secure validity was certainly unique. Customarily, validity is secured by comparing a test's score with another measurement of the same mental processes secured independently. The degree of relation is indicated by the amount of correlation which obtains between the two independent measures. In this case the validity would be indicated by the size of the coefficients computed between the Kuhlmann-Anderson test and (1) Stanford-Binet, (2) school marks, and (3) other group intelligence tests which have been tried out before. But these authors objected to each procedure in turn. They argued that (1) the Stanford-Binet test is an individual test and yields a score which is hard to compare with the group-test scores; (2) school

marks are a mixture of intelligence, interest, and teachers' whims and hence coefficients would be ambiguous to say the least, and (3) other group tests have used these just-criticized techniques to secure their validity and hence cannot be depended upon. These authors prefer to base their validity on the discriminative capacity of the test which they define as "the ability to make fine discrimination between small increments of mental development."¹ This means, for example, that there would be a sharp increase in percentage passing from, let us say, the seventh to the eighth years. There might be 40 per cent of the 7-year-olds who passed the test while 70 per cent of the 8-year-olds passed the same test. After all, school achievement, estimates of intelligence, and individual tests are as good indicators of intelligence as we have and should be utilized even though their weakness is recognized. The failure to compute these measures of validity weakens the test. How valid the test is cannot be determined. Such a successful test undoubtedly has high validity, but it was secured from the rich experiences of the authors and from the test's powers of discrimination.

Reliability

Kuhlmann and Anderson were also opposed to securing reliability in the usual way. Ordinarily the degree of reliability is indicated by a coefficient of correlated computed (1) between successive givings of the same test, (2) between successive givings of two forms of a test, (3) by the even-odd technique whereby the odd scores are correlated with the evens and then an estimate is made as to what the r would have been had the test been twice as long, or (4) by the application of the Kuder-Richardson formula (see page 29). But Kuhlmann and Anderson would have none of these. The variations in scores which were due to the change in the subject and not in the test would lead, they said, to a false interpretation, for the differences would appear to be in the test when it was really in the subject. They hold that the main cause of variations in scores is the shifts in interest and effort. These shifts are mostly caused by a failure of the tests to provide the right amount of difficulty for each subject. Since the Kuhlmann-Anderson tests are so well adjusted to the various ages, they argue, the effort is steady and the variation from one test to another is at a minimum. But here again *how* reliable the test is has not been determined. There is also a distinct advantage in having a test as reliable as possible under the best conditions of cooperation among the subjects and a definite interest in the test itself. If, then, there were variations in scores from one test to the next, thereby causing a reduction in correlation, we could know the part which variation in the subject beyond the normal played.

¹*Instruction manual*, p. 8. Educational Test Bureau, Minneapolis, Minn.

The difficulty with these simpler methods of computing reliability and validity arises in the fact that they are not quantitative. We want to know *how* reliable a test is, not *whether* it is reliable. We know the latter before we begin doing any calculation. If the reliability of one test calculated from a representative age group is .85 and that of another is .95, the second may definitely be used for individual diagnosis while the first may not be so used. Furthermore, if a group test correlates .70 with the Stanford-Binet and .55 with school marks, it is definitely to be preferred to one whose correlation is .50 with Stanford-Binet and .40 with school marks.

One other weakness in standardization appears. There is only one form. Two forms of a test are of real use when (1) it is suspected that the test has been spoiled, or (2) it is wished to prevent cheating by staggering the tests, or (3) it is desired to have an unusually reliable score by combining those of the two forms.

With all these shortcomings Kuhlmann-Anderson tests have been broadly and satisfactorily used. One competent user of tests, for example, says that he "has used the tests with entire satisfaction and considers them the most outstanding group scale available for use in the public schools."¹ One great advantage of this scale is that it does not reflect as much as some other group tests the results of teaching. As evidence, one may mention that four of the 10 tests used for grade 5 are not dependent either upon reading or upon other verbal relations.

PRIMARY MENTAL ABILITIES

In the development of explanations of intelligence Spearman showed that if the tetrad equations came out zero there were two components of intelligence, factors *g* and *s*. As studies continued in this area of intelligence it became clearer that in many cases of correlations among several tests the conditions of the two-factor theory were not satisfied. Other factors appeared whose clusters of tests correlated more closely with each other than with factor *g*. Spearman then introduced four or five group factors in addition to his factors *g* and *s*. These were thought of as supplementary.

The movement for factor analysis (led by such men as Thompson in England, and Thurstone) approached the matter in a different way. To them, intelligence could not be accounted for by a single dominant factor *g* but needed *several* coordinate factors to account for all the relations which exist in a large battery of tests. Among these American leaders Thurstone has not only worked out the theory and mathematics involved in factor analysis but has, with the support of Thelma Gwinn

¹ Turney, Austin H., *The 1938 Mental Measurements Yearbook* (Oscar K. Buros, ed.), p. 104. New Brunswick, N.J.: Rutgers University Press.

Thurstone, worked out tests which tap these factors which were theoretically independent or uncorrelated. To the Thurstones, intelligence is not a single entity which may be represented by an I.Q. or *g* but consists of seven or eight factors. For each of these factors tests have been constructed. There is, for example, one test all of whose items will measure speed of perception; another containing only items relating to memory; and still another made up of definitions of words, the *V* or verbal test.

When these factors were presented in a practical test suitable for a certain range of testing it was discovered that they did correlate positively with each other. To be sure these coefficients were not as high as in some other batteries, but they still were present. Here is a table of correlations from the *Examiner's Manual* of 1948.¹

| | <i>V</i> | <i>P</i> | <i>Q</i> | <i>Mo</i> |
|-----------|----------|----------|----------|-----------|
| <i>V</i> | | | | |
| <i>P</i> | .60 | | | |
| <i>Q</i> | .67 | .56 | | |
| <i>Mo</i> | .47 | .52 | .54 | |
| <i>S</i> | .55 | .61 | .56 | .46 |

V = Verbal meaning

P = Perception

Q = Quantitative

Mo = Motor

S = Space ("ability to visualize and to think about objects in two or three dimensions")

However, it must be said in fairness that the *interrelations* between the test factors decrease with age until at the college level the interrelations are in the order of .30 and not in the order of .50 as in the present case.

Because there are several components of intelligence, as the Thurstones believe, the general term "intelligence" need not be used. The stigma of the low I.Q. is in this manner averted and the subject's score may be shown him with impunity. For facilitating the pupil's understanding of his position on each ability, a PMA Profile Sheet is provided. The implications of his scores are printed on the back of the profile sheet to aid him in interpreting his own scores. Figure 31 shows Johnny Jones's scores on five primary mental abilities. Note Johnny's good scores on *V* and *P*, both of which are related to reading. Note that a mental age may be computed by combining the components.²

¹ Thurstone, Thelma Gwinn, and L. L. Thurstone, *Examiner's Manual for the SRA Primary Mental Abilities*, p. 7. Chicago: Science Research Associates, 1948.

² *Ibid.*, p. 12.

It is also claimed that differences in scores on these various factors are of great help for guidance. For example, there is a high correlation between scores on verbal meaning and perception on the one hand and readiness to read on the other. In like manner, the quantitative score gives a good idea of a child's possibilities in arithmetic.

Name JOHNNY JONES Sex Boy

School Community Elementary Date of Test 1948 9 4

Grade First Birth Date 1942 10 15

Room 101 Age 5 10 20

| AGE SCORES | Years Months | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|--------------------|-----------------|------------|------------|------------|----------------------|---------------------------|----------------------------|----------------------------|
| | | 2 4 6 8 10 | 2 4 6 8 10 | 2 4 6 8 10 | 2 4 6 8 10 | 2 4 6 8 10 | 2 4 6 8 10 | 2 4 6 8 10 |
| VERBAL-MEANING | Raw Score 37 | | | | | 31 40 25 | 43 43 | 44 45 |
| PERCEPTUAL-SPEED | 21 | | | | 22 25 | 24 25 | 26 27 | |
| QUANTITATIVE | 15 | | | | 16 18 20 17 19 | 22 23 | 24 25 | |
| MOTOR | 30 | | | | 34 36 37 | 38 40 42 40 42 | 44 46 48 42 44 | 50 52 54 48 50 52 |
| SPACE | 14 | | | | 15 17 19 16 18 | 21 20 | 22 23 | 24 25 |
| AGE SCORES | Years Months | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| | | 2 4 6 8 10 | 2 4 6 8 10 | 2 4 6 8 10 | 2 4 6 8 10 | 2 4 6 8 10 | 2 4 6 8 10 | 2 4 6 8 10 |
| TOTAL (V P-Q-S) | 87 | | | | 94 96 98 90 92 94 | 100 102 104 98 100 102 | 106 108 110 104 106 108 | 112 114 116 110 112 114 |

FIG. 31. Johnny Jones's score on five primary mental abilities.

The reliabilities of the primary mental abilities when computed by the Spearman-Brown method, using 500 students in grade 10B, are:

| | |
|---------------------|-----|
| Verbal meaning..... | .92 |
| Space..... | .96 |
| Reasoning..... | .93 |
| Number..... | .89 |
| Word fluency..... | .90 |

Some of the claims for this series of tests have been substantiated statistically, but they (PMA) have not had such wide use as many of the other tests that have been on the market longer. Here are the results of investigations, not before published.¹ Relationships of each of the five primary mental abilities (PMA) to marks on subjects studied in high school were computed. Of particular interest are the coefficients computed between *V* (verbal meaning) and marks in English. Consider

¹ Moody, Caesar B., *Analysis of SRA Primary Mental Abilities of High School Pupils*, doctoral dissertation, University of North Carolina, 1951.

the fact that the test of verbal meaning takes only 4 minutes of working time and yet achieves coefficients between .50 and .72 with marks in English, reading, civics, and United States history. Furthermore, this same *V* correlates .76 with marks in general business and .66 with home economics. Even in elementary science and biology the coefficients are substantial. A study of Moody's tables indicates other interesting relations with school marks. Space has no high relation with school marks. Reasoning's highest correlations are with English III (.66), United States history (.67), and typing (.65). Number shows substantial coefficients with United States history, typing, and general mathematics. In many cases the single primary mental ability shows a closer relationship with school marks than when the five are combined (*T*).

If other studies agree with this one, no general intelligence test will surpass in usefulness for high school testing and guidance the SRA Primary Mental Abilities, intermediate, ages 11 to 17.

INTELLIGENCE TESTS FOR VARIOUS LEVELS

KINDERGARTEN AND BEGINNING FIRST GRADE

At the end of this section appears a selected list of group tests of intelligence suitable for the kindergarten and the beginning first grade. Below the level of the kindergarten it is almost impossible to administer a group test satisfactorily, and even at the level of kindergarten and the beginning first grade there are difficulties enough. The attention of children of this age shifts easily from one object to another. They are not yet accustomed to work on a topic more than a few minutes. Negativism may appear at almost any time and express itself in a downright refusal to cooperate. Finally, great variation appears in children's efforts unless they are genuinely interested in the materials utilized in the test. Test makers have done their best to construct items in such a manner as to keep attention on the problem assigned, to avoid wandering of attention, and to ensure steady effort. They have utilized attractive pictures to be described or to discover what was wrong or missing in them, simple and more complicated drawings to be copied, and pictures of objects to be counted. No written materials can possibly be used. Probably not more than 10 beginning first graders should be tested at one sitting. The tester should see to it that each little subject gets his own test blank, that they all have the right place before beginning, that they do not simply copy from each other, and that any child's attention with a propensity to wander be brought back immediately to the problem at hand. More than at any other age good results depend upon the cleverness of the tester in manipulating the testing process so as to get the best effort possible from each subject.

At no other age level is there a greater need for an accurate appraisal of a child's intelligence than in the first grade. Such an appraisal enters heavily into any decision to begin the more formal work of the first grade (reading, numbers, writing, etc.). On the contrary, the reliabilities of the tests suitable for such testing are lower than those of the upper grades.

Some good group tests for kindergarten and beginning first grade are (1) Pintner-Cunningham Primary Mental Test, revised, kindergarten to grade 2; (2) Kuhlmann-Anderson Intelligence Tests, grade 1, first semester; (3) Detroit Beginning First Grade Intelligence, revised 1935; (4) Goodenough Intelligence Test, kindergarten to grade 3; (5) California Test of Mental Maturity, kindergarten and grade 1, 1943 preprimary battery; (6) SRA Primary Mental Abilities, PMA, ages 5 to 7.

GRADES 1 THROUGH 3

Materials for tests of these school grades show a definite change from concrete, pictorial materials to the use of language and number. The language in the first instant is oral; the answer being registered in a picture. In the second case written language is used both in the situation and in the response. Let us take analogies as an illustration (Pintner-Durost, Scale I-A). The situation is given orally: "robin: worm." The subject then must find in pictures the same relation. There are four pictures: a cat, a dog, a cat at a piano, and a mouse. Robin: worm = cat: mouse. When the relation is a written verbal one, the problem is, a "clock: time = thermometer: mercury—zero—temperature." The clock is to time as the thermometer is to temperature. In the opposites test a similar condition holds. For illustration the question is given orally "Mark the picture that means the opposite of asleep." The answer is contained in three pictures: (1) a bed, (2) a child evidently asleep in bed, and (3) a child sitting up reading.

The other tests of this series in one form contains all answers in pictures while in the other form, the problem and the solution are written.

Test forms found most satisfactory at this level are the ones that have been tried out on numerous occasions and have proved their worth. Opposites and analogies have already been mentioned. Arithmetic reasoning holds its own as a test form. Vocabulary tests, both oral and written, remain good. Among the younger children the copying or completion of drawings or the recognition of a drawing among others closely similar have been used. Tests of classification deserve special mention. These tests demand that the subject see likeness between items apparently different and then mark out another item really different from the other four. Altogether in the four recent tests especially suitable at this level there are 25 different types of test forms which have been included

- "No. 1. Mark the picture that means the opposite of straight.
 "No. 2. Mark the picture that means the opposite of high.
 "No. 3. Mark the picture that means the opposite of rough.
 "No. 4. Mark the picture that means the opposite of push.



FIG. 32. Pintner-Durost Elementary Test, Test 4, opposites: picture content.

in the battery (1) because they show a sharp rise in percentage passing from one age to the next higher one or from one grade to the next higher one, and (2) because they correlate substantially with the total test. Nearly all the tests require the perception of relations to pass them successfully, although a few simply require keen observation and memory.

TEST 4. OPPOSITES

In each line mark the word that means just the *opposite* of the first word.

| | | | | |
|-------------|---|--|--|--|
| A. black — | <u>dark</u> <input type="radio"/> | <u>light</u> <input type="radio"/> | <u>white</u> <input type="radio"/> | <u>night</u> <input type="radio"/> |
| B. down — | <u>below</u> <input type="radio"/> | <u>high</u> <input type="radio"/> | <u>top</u> <input type="radio"/> | <u>up</u> <input type="radio"/> |
| <hr/> | | | | |
| 1. fast — | <u>slow</u> <input type="radio"/> | <u>careful</u> <input type="radio"/> | <u>quick</u> <input type="radio"/> | <u>driving</u> <input type="radio"/> |
| 2. hard — | <u>soft</u> <input type="radio"/> | <u>kind</u> <input type="radio"/> | <u>rough</u> <input type="radio"/> | <u>strong</u> <input type="radio"/> |
| 3. clean — | <u>dirty</u> <input type="radio"/> | <u>spotless</u> <input type="radio"/> | <u>noise</u> <input type="radio"/> | <u>house</u> <input type="radio"/> |
| 4. strong — | <u>big</u> <input type="radio"/> | <u>weak</u> <input type="radio"/> | <u>men</u> <input type="radio"/> | <u>small</u> <input type="radio"/> |
| 5. young — | <u>antique</u> <input type="radio"/> | <u>youth</u> <input type="radio"/> | <u>little</u> <input type="radio"/> | <u>old</u> <input type="radio"/> |
| 6. quiet — | <u>cool</u> <input type="radio"/> | <u>still</u> <input type="radio"/> | <u>soft</u> <input type="radio"/> | <u>noisy</u> <input type="radio"/> |
| 7. find — | <u>keep</u> <input type="radio"/> | <u>drop</u> <input type="radio"/> | <u>lose</u> <input type="radio"/> | <u>discover</u> <input type="radio"/> |

FIG. 33. Pintner-Durost Elementary Test, Test 4, opposites: reading content.

Some good tests for grades 1 through 3 are (1) the Pintner-Durost Elementary Test,¹ suitable for last half of grade 2, grade 3, and first half of grade 4 (Figs. 32 and 33); (2) the California Test of Mental Maturity, grades 1-3; (3) the Kuhlmann-Anderson Intelligence Tests (in separate booklets), grade 1 (second semester), grade 2, and grade 3; (4) the Otis Quick Scoring Mental Ability Tests, Alpha Test, grades 1 to 4; and (5) the SRA Primary Mental Abilities, PMA, ages 7 to 11. From the Cali-

¹ Items by permission of World Book Company, Yonkers, N.Y.

| B. Spatial Relationships . . 32 | |
|--|--|
| 6. Sensing Right and Left . . 10 | 4 6 8 10 12 14 16 18 20 22 23 24 25 26 27 28 29 30 31 32 |
| 7. Manipulation of Areas . . 12 | 2 3 4 5 6 7 8 9 10 11 12 |
| 8. Foresight in Spatial Sit'ns . . 10 | 1 2 3 4 5 6 7 8 9 10 |
| C. Reasoning 48 | |
| 9. Opposites 12 | 5 10 15 20 25 30 35 40 45 50 55 60 65 70 72 |
| 10. Similarities 12 | 1 2 3 4 5 6 7 8 9 10 11 12 |
| 11. Analogies 12 | 1 2 3 4 5 6 7 8 9 10 11 12 |
| 12. Number Concepts . . 12 | 1 2 3 4 5 6 7 8 9 10 11 12 |
| G. Non-Language Factors . 100 | |
| H. Chronological Age . . . | |
| I. Actual Grade Placement . (Grade pupil is in) | |
| <div> <div>48 60 72 84 96 108 120 132 144 156 168</div> <div>4.0 5.0 6.0 7.0 8.0 9.0 10.0 11.0 12.0 13.0 14.0</div> </div> | |
| Mental Age | Yr. Mo. |

SUMMARY OF DATA

G. Non-Language Factors

FIG. 34. Form for representing each subject's score, California Test of Mental Maturity.

fornia Test one may secure language and nonlanguage M.A.s as well as an M.A. based on the total test. It attempts to divide total intelligence into (1) memory, (2) spatial relationships, (3) reasoning, and (4) vocabulary. Also a dichotomous classification is made into language and nonlanguage tests (Fig. 34).

GRADES 4 THROUGH 8

The intelligence tests suitable for these grades employ many of the same test forms that were used in grades 1 to 3. The items of these test forms or subtests are made more difficult by using more complex materials and by making all the choices more plausible ones. The test forms occurring most frequently are opposites and number completion, followed closely by logical selection, classification, analogies, and arithmetic reasoning.

The giving of opposites to words permits of almost infinite range in difficulty. Samples of opposites picked from tests suitable for these grades are:¹

Find—1. penny 2. get 3. keep 4. lost 5. lose [Pintner]
Which word means the opposite of humility?

1. joy 2. pride 3. dry 4. funny 5. recklessness [Otis]
Find both—tennis easy punish lesson nice reward
[Kuhlmann-Anderson]

Number completion appears in two forms. In one of them the problem is to find what number of the series is wrong:

1—2—4—8—14—16—32 [California]
3—7—11—13—15—19 [Kuhlmann-Anderson]

In the other, the sequence of numbers is to be completed:

5—9—13—17—21—25— (a) 30 (b) 28 (c) 27 (d) 29 (e) 26
[Pintner]
 $\frac{1}{2}$ — $\frac{1}{4}$ — $\frac{1}{8}$ —1—3—9— (a) 12 (b) 27 (c) 15 (d) 18 (e) 32
[Pintner]

Test forms of logical selection, classification, analogies and arithmetic reasoning have been widely used. In logical selection the question usually is, "What do these things always have?" or, expressed in another way, "What are these things never without?"

¹ In this chapter permission for the use of the Pintner items was received from the World Book Company, Yonkers, N.Y.; for that of the California items, from the California Test Bureau, Los Angeles, Calif.; for that of the Kuhlmann-Anderson items, from the Educational Test Bureau, Minneapolis, Minn.

River—(1) fishes (2) boats (3) banks (4) bridge (5) ferry
[Pintner]
Squirrel—(1) nuts (2) fur (3) tail (4) cage (5) tree
[two things—Kuhlmann-Anderson]

In classification the problem is to discover in what respects four of the items are alike and one is different and to cross out the one that is not like the others.

(1) diamond (2) gold ~~(3)~~ ruby (4) iron (5) platinum [Pintner]
(1) general ~~(2)~~ ensign (3) major (4) colonel (5) captain
[Kuhlmann-Anderson]

The test form of analogies has been with us from the time of the first test in group testing and is still highly regarded. In analogies one discovers a relation between two items and then applies that discovered relationship to the solution of the problem.¹

Body: Food: Engine (1) wheels (2) motion (3) smoke (4) fire
(5) fuel [Pintner]
A lamp is to a light as (?) is to a breeze—(1) a fan (2) bright (3) a sailboat
(4) a window (5) blow [Otis]

Another old war horse in test construction is arithmetic reasoning. It has weathered the criticisms of being a special ability or of being too much like school because it correlates highly with the total test and because it is passed by a larger percentage of subjects at each increasing age level.

The sum of two numbers is 100. One of the numbers is 35. What is the other number?
(a) 135 (b) 3500 (c) $2\frac{5}{6}$ (d) 65 (e) 30. [Pintner]

In a field meet, 20 events were listed for the day. Pupils from your school won 60 per cent of the events. How many events did you lose? (1) 4 (2) 3 (3) 8 (4) 12 [California]

What is the number $\frac{1}{3}$ of which is $\frac{5}{6}$ of 18? [Kuhlmann-Anderson]

It is immediately apparent that the six texts most frequently used in test construction at this level of intelligence are for the most part expressions of relationships between facts well known by the subject. Now and then an error appears because of a lack of experience with the original data, but this is not the rule. High scores are secured by those who are able to perceive relationships among words, numbers, or visual areas. The other tests which are now listed are for the most part designed to test the subjects' capacity to discover relations more or less clearly apparent.

¹ Permission to use items from the Otis test cause from the World Book Company, Yonkers, N.Y.

The third group of test forms is composed of vocabulary, best answer or logical reasoning, substitution, memory, and similarities. In the first Stanford Revision of the Binet-Simon tests, Terman placed great emphasis upon his vocabulary test. He thought it as good as two or three ordinary test items. While not quite as high a position would be given it today, it still is regarded as a useful test.

| | | | | |
|-------------------|-------------|--------------|------------|---------------------|
| refuse—(1) object | (2) accept | (3) delay | (4) reject | (5) value [Pintner] |
| ballet—(1) feast | (2) banquet | (3) carnival | (4) ball | (5) dance |
| | | | | [Pintner] |
| dispute—1 disturb | 2 question | 3 subdue | 4 disguise | [California] |

The best-answer test, too, appeared in the original Army Alpha. In this test the subject selects the best answer out of three or four plausible answers.

"Drop by drop the lake is drained" means:

- (1) Every man wishes water for his own well.
- (2) It is never too late to mend.
- (3) Drowning men will catch at a straw.
- (4) All's well that ends well.
- (5) Many little strokes fell great oaks.

[Pintner]

Either the sun moves around the earth or the earth moves around the sun. But the sun does not move around the earth. Therefore

- (1) the earth moves around the moon.
- (2) the earth moves around the sun.
- (3) the sun is larger than the earth.

[California]

Another test form that has age on its side is the substitution test. It became popular perhaps because it reflected easily and directly the results of learning. Since intelligence was in some quarters defined as the "capacity to learn," this test fitted directly into that definition. One may have a key such as

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| A | E | U | B | D | G | C | F | H |

and then be asked to write 416, 1632, or 425134 using this key.

A test of memory may be constructed by giving a series of words in pairs, then giving the first member of the pair and asking for the second one. One reads first: "wind—tree, nine—four, sleep—bed, river—fish." He is then given the word "wind" and asked to find one picture from four pictures which completes the pair. Another procedure is to read aloud to the group being tested a story, then 15 or 20 minutes later ask questions about the story.

The last test form to be discussed in this group is that of similarities. Here one discovers in what respects two or three things are alike and

picks out of several others the one which is similar to the first two or three. This procedure may be carried out either with words or pictures.

Which of the five things below is most like these three: a tent, a flag, a sail?

1 a shoe 2 a ship 3 a staff 4 a towel 5 a rope
() House () Cave () Barn () Hotel () Store () Castle
[“Mark three that are alike.”—Kuhlmann-Anderson]

There are many other forms which have been successfully used. Right and left, mazes, anagrams, mixed sentences, recognizing visual units in concrete patterns, range of information, dividing visual figures, hard directions, using alphabet, giving the genus of a named species, and several others. Of all these, only four will be described. In hard directions, the alphabet may appear at the top of the page and then such questions as “The first letter to the left of the 10th letter is—?” Anagrams have possibilities of great complications:

E—P—N—L—C—I. What is the word?

M—O—S—U—E

[Kuhlmann-Anderson]

The range-of-information test was used as a member of the original Army Alpha:

| | | | | |
|-----------------------|------------|------------|----------|--------------|
| Leghorn is a kind of: | 1. rabbit | 2. chicken | 3. cow | 4. horse |
| | 5. sheep | | | [California] |
| Veins are found in: | 1. flowers | 2. leaves | 3. seeds | 4. petals |
| | 5. roots | | | [California] |

Mixed sentences were used in the Stanford Revision of the Binet-Simon tests. It is a question of unscrambling sentences and then making some judgment about them

children room of the out ran six

[“Mark first and last word of corrected sentence.”]

who her lost girl pencil the another bought

[Kuhlmann-Anderson]

Suppose we consider all these successful test forms in the light of theory. One of the first theories set forth was the two-factor one. Spearman, as we have discovered, emphasized two factors in his explanation of intelligence, factor *g* and factor *s*. Factor *g* approaches very closely our usual term of general intelligence. Spearman then inquired into the characteristic of those tests that were heavily loaded with *g*. He found two principles of explanation: (1) education of relations, and (2) education of correlates. In the education of relations two items were set down and a relation discovered between them. *e.g.*, “often—seldom (same or opposite?).” In education of correlates one might give the word “often” and ask for its opposite, or analogies might be used as, “Sheep: mutton:: pig: (1) lamb (2) meat (3) pork (4) beef.” When we consider our statistically

successful tests in the light of these two principles of relations we find a remarkable number of them concerned with the perception of relations. Let us consider more minutely the six most successful test forms: opposites, number completion, logical selection, classification, analogies, and arithmetic reasoning. In each of these the perception of relations is the dominant characteristic. Indeed opposites and analogies are illustrations par excellence of the perception of relations. In number completion the relation between the numbers in a sequence must be discovered. In classification similarities between some members must be perceived in order to isolate the dissimilar one. And so it goes with logical selection, in which one chooses what an object always has, and arithmetic reasoning, in which relations must be comprehended in order to proceed to the proper solution of the problem. Nor is there a great deal of difference when the other forms are considered. Vocabulary, best answer, and substitution are pretty largely tests of the capacities to perceive relations. Spearman would say that those tests in which perceptions of relations either by education of relations or the education of correlates are heavily loaded with *g* are good tests of intelligence.

The following are good tests for grades 4 to 8: (1) Pintner General Ability Tests, Verbal Series, intermediate test, grades 5 to 8, (2) California Test of Mental Maturity, elementary series, grades 4 to 8, (3) Kuhlmann-Anderson Intelligence Tests, Test 1 for grade 4, Test 2 for grade 5, Test 3 for grade 6, Test 4 for grades 7 and 8; (4) Detroit Alpha Intelligence Test, grades 4 to 8, Form T; (5) SRA Primary Mental Abilities, PMA, ages 7 to 11 and also ages 11 to 17.

HIGH SCHOOL—GRADES 9 THROUGH 12

The same test forms already mentioned for grades 4 to 8 are also used in the high school. The relations expressed are more subtle and therefore more difficult to discern. Among the test forms found successful by nearly all test makers are analogies, arithmetic reasoning, opposites, vocabulary, and number sequences.

In analogies is this more difficult relation made clear. Illustrations are:¹

| | | | | | |
|-----------------------------------|----------|----------|---------|-------------|---|
| peace—happiness::war— | 1 sorrow | 2 fright | 3 death | 4 bellicose | |
| 5 trouble | | | | | [Pintner] |
| tree is to forest as person is to | 6 women | 7 couple | 8 human | | |
| 9 crowd | 10 men | | | | [Terman-McNemar] |
| Japanese | Japan | Russian | Dutch | Serbia | Spanish |
| | | | | | Holland |
| | | | | | [Pick out both relations—Kuhlmann-Anderson] |

¹ Permission for items from Terman-McNemar Test from the World Book Company, Yonkers, N.Y.

Arithmetic reasoning is so well known that only one illustration will be used:

If a boy can run at the rate of 6 feet in $\frac{1}{4}$ of a second, how far can he run in 10 seconds? [Otis]

Opposites are again used both in words and in pictures:

Obtuse—1 accessible 2 abstruse 3 acute 4 corpulent 5 agile
[Terman-McNemar]
Affinity 1 capillarity 2 consanguinity 3 gravitation 4 magnetism
5 repulsion [Pintner]

Vocabulary continues into the high school its usefulness as a test form:

diurnal—1 weekly 2 yearly 3 nightly 4 daily 5 monthly
[Pintner]
recumbent—1 cumbersome 2 curved 3 reclining 4 saving
[California]
curdle—1 coagulate 2 spoil 3 snuggle 4 condense 5 churn
[Terman-McNemar]

Number sequence is at the high school level as well as in the previous grades one of the most useful test forms:

$\frac{1}{4}$ $\frac{1}{8}$ $\frac{1}{16}$ $\frac{3}{16}$ $\frac{5}{8}$ — (a) $\frac{3}{8}$ (b) $\frac{3}{8}$ (c) $1\frac{5}{16}$ (d) $\frac{9}{4}$ (e) $1\frac{7}{8}$
[Pintner]
32 29 27 22 17 12 [cross out wrong number -Kuhlmann-Anderson]
60—55 51 49—40 37 [fill in gaps—California]

Along with these generally accepted test forms are others whose usefulness is unquestioned: best answer, logical selection, classification, disarranged sentences, hard directions, similarities, and memory. There is not a great deal of difference between best answers and logical selection:

“Better give a shilling than lend a half crown” means—

1. Better a penny than a copper.
2. Better give the wool than lend the whole sheep.
3. Give little to the big.
4. A shilling grows bigger with years.
5. A shilling will buy a crown.

[Pintner]

Notice the slight difference between the illustration above and those under logical selection. Here the problem is to discover what the thing always has:

A prism—1 triangle 2 parallelogram 3 glass 4 octagon 5 pentagon
[Pintner]
Compromise always involves: 6 respect 7 friendship 8 adjustment
9 law 10 violation [Terman-McNemar]

Classification involves the crossing out of a word or picture which does not belong with the others:

| | | | | | |
|-------------|--------|------------|----------|-------------|------------------|
| 1 trapezoid | 2 cube | 3 triangle | 4 square | 5 rectangle | [Pintner] |
| 6 large | 7 tall | 8 high | 9 short | 10 low | [Terman-McNemar] |

Disarranged sentences are also useful at these upper levels:

| | |
|--|---------------------|
| Mark first and last word in correctly arranged sentence—children room of the out | |
| ran six | [Kuhlmann-Anderson] |
| period of a this close at the put sentence | [Miller] |

Hard directions were used in our first group test:

(Alphabet printed at top of page) Write the letter which follows the letter which comes next after C in the alphabet. [Otis]

Think what year this is then write here _____ the digits in the reverse order. Put in the correct signs in this example $12 \quad 2 \quad 6 = 30$ [Kuhlmann-Anderson]

In similarities the likeness of two or three words or pictures are discovered; then the word or picture agreeing with this likeness is marked.

| | |
|--|--------------|
| large, red, good—heavy, size, color, apple, very | [Otis] |
| (In pictures) hammer, anvil, nut to fit a bolt—electric light bulb, glass jar, water tap, and rolling pin. | [California] |

The final test form in this group is a test on delayed memory. This test may be for immediate or delayed memory. In one form, a passage is read; then questions about it follow immediately or in other batteries after 25 to 30 minutes. In another, words are read in pairs; then the first word of the pair is given and the idea of the second word is found among 3 or 4 pictures, *e.g.*, safety—key; graceful—swan; clear—ice; power—boat; hungry—lion; resting—acorn; base—triangle; circles—spring, danger—sailor. After these pairs are read, the word *safety* is given and the subject will select *key* from among other pictures if his memory is good.

In addition to these tests listed above, whose use is widespread, occur many types of test forms used only in one test battery.

From these illustrations and from the study of whole battery of tests it is clear that good tests of intelligence can be built out of materials known to the vast majority of students. In most cases the successful passing of the tests involves the perception of more or less subtle relations existing between materials already experienced.

Suitable tests for this level (grades 9 to 12) are (1) Pintner General Ability Tests, advanced test, grades 9 to 12; (2) Terman-McNemar Test of Mental Ability; (3) California Test of Mental Maturity, advanced series, grades 7 to 12; (4) Kuhlmann-Anderson Intelligence Tests,

grades 9 to 12; (5) Otis Group Intelligence Scale, advanced examination, grades 7 to 12 (self-administering); (6) SRA Primary Mental Abilities, PMA, ages 11 to 17.

General Characteristics of Test Forms

From the consideration of the pages of description of tests in this chapter one can get a fair understanding of the types of test forms which test makers have found useful. In general, they all are passed by an increasing percentage of subjects with increasing age and all of them correlate well with the total test. Relations are expressed in a variety of test forms and in several media. Visual forms and word forms are by far the most frequent media in which test items appear. In some tests, analogies or opposites, for example, may be given first with pictures (nonlanguage) and then with words (language). Pictures are widely used with young students before they learn to read and with those who by some environmental condition are handicapped in their vocabulary and reading development. The same test forms appear at many levels of development. The increasing difficulty of these instruments is related to the greater subtlety of relationship between the facts or words, the increasing rareness of the words or other materials, and the increasing degree of similarity of the several answers from which one must be selected. With some exceptions, tests are dependent upon the education of relations and the education of correlates for the successful answers. It is important to observe that all tests use largely the same forms and that the superiority of one test over another depends upon the careful checking of each item and the ingenuity in selecting items which challenge directly the mental functions desired or, more precisely, that correlate best with the desired criteria.

USES OF INTELLIGENCE TESTS

At the very beginning of the child's entrance into the formal school work of the first grade, intelligence tests are of primary use. Whatever else these tests measure, they measure something that is related to the capacity to learn to read, write, and figure. The law of the land usually requires that a child enter school when he is 6 years of age. Note that this requirement is in terms of chronological years, not mental years. But children of the same chronological age differ greatly in mental age.

For example, one student comments as follows upon McNemar's study of 2,106 subjects in grades 1 to 12 tested by the Terman-Merrill Revision:¹

¹ Cook, Walter W., in *Educational Measurement* (E. F. Lindquist, ed.), pp. 9-10. Washington, D.C.: American Council on Education, 1951.

One may conclude from these and other data presented in this study that in a typical school: (1) the first-grade teacher will find that 2 per cent of the pupils have mental ages of less than four years and that 2 per cent will have mental ages of more than eight years; (2) the sixth-grade teacher will find that 2 per cent of the pupils have mental ages of less than eight years and that 2 per cent will have mental ages of more than sixteen years; (3) the high school teacher will find a range of from eight to ten years in mental age at each grade level; and (4) these conditions will be found to exist whether the school enforces strict policies of promotion and failure or promotes entirely on the basis of chronological age.

How great this variation is has also been clearly indicated in the study of 4,393 first-grade children.¹ Not all of them were 6 years of age, since they varied in age from 5-4 to 13-2, but the range of Mental Ages in this group was enormous. They ranged in M.A. from 2-10 to 10-2. Children need an M.A. of 6 years or thereabouts to learn efficiently and happily the work of the first grade. Evidence for the necessity of an M.A. of 6 for successfully doing work in the first grade appeared some years ago (1922). In this instance, dealing with 277 first graders, 81 per cent of those who had an M.A. of 6 years were promoted from 1B to 1A, 59 per cent of those whose M.A.s ranged between 5-8 and 6-0, and none of those whose M.A.s were below 5-8.² Recent studies show that children whose M.A.s are as low as 5-0 may be taught to read if the materials are carefully selected for this age. But the going is most certainly slow and hazardous.

Webb and Shotwell³ present interesting illustrations of the uses of tests with children of superior ability, with those slightly below normal, and with the definitely feeble-minded whose individual needs were met by careful planning. In one case a 5-year-old girl with an M.A. of 8-0 was advised to go to a private school, where she led her group of normal 6-year-olds and continued her good work into the second grade. Such a program of study could not have been undertaken with a socially backward, physically retarded child. Other cases are presented where parents were definitely advised to keep their children in kindergarten another year because their mental levels were clearly inadequate for success in the ordinary work of the first grade.

¹ Dickson, V. E., *Mental Tests and the Classroom Teacher*, pp. 96-97. Yonkers, N.Y.: World Book Company, 1923.

² Davis, H., "Intelligence Tests in Public Schools in Jackson," *Twenty-first Yearbook of the National Society for the Study of Education*, Chap. III, pp. 131-142. Bloomington, Ill.: Public School Publishing Company, 1922.

³ Webb, L. W., and Anna Markt Shotwell, *Testing in the Elementary School*, pp. 114-116. New York: Rinehart & Company, 1939.

Enough has been said to make it abundantly clear that the scores secured from such tests as have been previously described may serve a practical function in determining the approximate time for a first-grade teacher to begin formal instruction in reading, writing, etc. One remembers, of course, that factors other than intelligence enter into reading readiness. Intelligent parents who read to children, who answer their questions, who take them walking and tell them stories undoubtedly raise the children's vocabulary level and thus make them more ready to learn to read. But even here intelligence is related to the number of words learned and the understandings acquired.

In the second place, intelligence tests are helpful in guiding students into those courses where they have some likelihood of being successful. It is now well known that certain school subjects are much more closely correlated with intelligence tests than are others. This means that those who score high on intelligence tests also do very well on these subjects; those who have medium scores on the intelligence tests get about average marks on these subjects, and finally the lower third have a great deal of difficulty with these subjects. In the elementary school, composition, reading for understanding, dictation, and arithmetic problems usually have coefficients with intelligence tests of .5 to .6. In the high school, subjects such as mathematics, Latin, and English composition are highly dependent upon intelligence. Professor Thorndike, who gave especial attention to this problem, thought that the correlation between algebra and intelligence in the high school would ordinarily be in the neighborhood of .70,¹ although a relation of .45 and .50 would more nearly represent what is usually found.

Let us look for a moment at (1) the intelligence of those who elect various high school subjects, and (2) the intelligence required of those who pass the courses. What levels of intelligence do those persons possess who elect solid geometry and trigonometry? According to one investigator,² more than three-fourths of high school students electing solid geometry and trigonometry come from the upper fourth in intelligence and less than 10 per cent from the lower fourth. Latin, natural science, Spanish, and French also drew heavily from those with high intelligence. In the second place, the median intelligence quotients for those boys passing high school subjects also varied widely. The highest I.Q.s were possessed by those who passed Latin, followed next by the I.Q.s of those who passed ancient history and algebra.³

¹ Thorndike, E. L., *The Psychology of Algebra*. New York: The Macmillan Company, 1923.

² Powers, S. R., "Intelligence as a Factor in the Election of High School Subjects," *School Review* (1922) 30:452-455.

³ Madsen, I. N., "The Contribution of Intelligence Tests to Educational Guidance in High School," *School Review* (1922) 30:692-701.

At various levels of education, the story is repeated. In college, most difficult and most exacting in intelligence are mathematics, the natural sciences, and the foreign languages. In the elementary school, on the contrary, handwork, drawing, and handwriting correlate very low with intelligence scores. At the high school level, manual training, mechanical arts, and domestic arts have low correlations with intelligence. The I.Q.s of those who pass them are measurably lower, and the majority of those electing them are below average in verbal intelligence. Commercial subjects at the high school level are elected largely by those students who are somewhat below the average of other students in intelligence.

These facts are convincing evidence for the use of tests in educational guidance. Intelligence scores can be used to advise pupils and students concerning the subjects they may take. It seems clear that the guidance proffered to those of superior intelligence will depend more upon their interests or upon the vocation toward which they are looking. For those students falling below average in intelligence and more especially for those whose I.Q.s are between 80 and 90, problems of choosing subjects with some possibility of success loom very large. When subjects are selected that are not too much loaded with intellectual content, the number of students continued in school is greatly increased. Sometimes those who are below 90 I.Q. take foreign language, for example, with not very satisfactory results.

In connection with a study by Oscar H. Werner¹ the author has written²

One of the findings of this study has such a general implication that we may be permitted to lift it out of its context and give it a more general setting. This finding relates to the greater improvement of those with higher intelligence who study modern foreign languages. When the students were divided into three groups: (1) the low group, I.Q.'s 85 to 89, (2) the medium group, those with I.Q.'s from 95 to 104, and (3) the high group with I.Q.'s of 110 to 114, then the higher the I.Q. group the greater the improvement in desirable English abilities. Those of low intelligence seem really to have become confused, and to have done worse on English abilities than they had done before taking up the study of modern foreign languages. They showed losses in the English tests in five out of six cases. Even those of average capacity lost more than they gained.

¹ Werner, Oscar H., "The Influence of the Study of Modern Foreign Languages on the Development of Desirable Abilities in English," *Studies in Modern Foreign Language Teaching*, pp. 99-145, Publications of the American and Canadian Committees on Modern Languages. New York: The Macmillan Company, 1930.

² Jordan, A. M., *Educational Psychology*, 3d ed., pp. 292-293. New York: Henry Holt and Company, Inc., 1942. By permission.

But the pupils of high I.Q.'s made substantial gains in all the tests of English abilities with the exception of tests of punctuation and sentence structure. It is they who really understand a foreign language and have the mental capacity to see relations and contrasts between the two languages which enable them to make large improvements in scores in grammar, language usage, and reading. It can almost be made a universal that no student whose I.Q. is below 90 should be allowed to register for a modern foreign language.

RESULTS OF EDUCATIONAL GUIDANCE

As was stated on page 404 the results of such careful consideration of individual differences and the adaptation of courses to them produce a visible, measurable effect. For purposes of contrast we shall introduce the usual elimination of students when there is no definite program of guidance and contrast with these conditions the results when guidance has been effectively done.

TABLE 13.* RELATIONSHIP OF SCHOOL SUCCESS TO BINET INTELLIGENCE QUOTIENT (131 high school pupils tested in 1916 and 1917 and followed up for 6 or 7 years)

| I.Q. on Stanford-Binet scale | Number of cases in each group | Completed 4-year high school course | | Left high school to go to work | |
|-------------------------------------|-------------------------------|-------------------------------------|----------|--------------------------------|----------|
| | | Number | Per cent | Number | Per cent |
| 125 or over (very superior) | 19 | 19 | 100 | 0 | 0 |
| 115-125 (superior) | 27 | 26 | 96 | 1 | 4 |
| 105-114 (above average) | 24 | 20 | 83 | 4 | 17 |
| 95-104 (average) | 36 | 27 | 75 | 9 | 25 |
| 85-94 (below average) | 22 | 9 | 40 | 13 | 60 |
| 78-84 (dull) | 3 | 0 | 0 | 3 | 100 |
| Totals | 131 | 101 | 77 | 30 | 23 |

* Proctor, W. M., *Educational and Vocational Guidance*, Table II, p. 31, Riverside Textbooks in Education. Boston: Houghton Mifflin Company, 1925.

Table 13 offers evidence that the bright continue in school and that the dull are eliminated. Out of 19 students with an I.Q. of 125 or better, all finished high school. Contrast this record with that of the students with an I.Q. of 85 and below. Only 9 of these 25 were able to finish high school, or about one-third of the total. A similar story is shown in Table 14. In this table the number of years concerned is only 2 instead of the 4 studied in the high school. But even in 2 years the trend is inescapable. The correlation between intelligence scores and length of stay in college

TABLE 14.* THE RELATIONSHIP BETWEEN SCORES ON OTIS TEST AND THE CONTINUATION OF STUDENTS IN COLLEGE
(Study covers a period of 2 years)

| I.Q. derived from Otis Tests | Total number of students | Those remaining 2 years | | Leaving before 2 years | |
|------------------------------|--------------------------|-------------------------|----------|------------------------|----------|
| | | Number | Per cent | Number | Per cent |
| 115-124 (superior)..... | 158 | 115 | 72 | 43 | 28 |
| 105-114 (above average)..... | 247 | 154 | 62 | 93 | 38 |
| 95-104 (average)..... | 103 | 60 | 57 | 43 | 43 |
| 85-94 (below average)..... | 43 | 18 | 42 | 25 | 58 |
| 75-84 (dull)..... | 11 | 2 | 18 | 9 | 82 |
| Total..... | 562 | 349 | 62 | 213 | 38 |

* Jordan, A. M., *Educational Psychology*, 3d ed., p. 520. New York: Henry Holt and Company, Inc., 1925.

is substantial. In Table 14 note how the percentages in column 3 decrease from 72 to 18 as the intelligence of the groups decreases.

In these two studies we have clearly demonstrated what happens to students when there is no program of guidance. Another result of allowing students to drift into courses for which they are intellectually unprepared is to reduce the level of the work of the college preparatory courses. As a consequence, those who are really capable of first-class intellectual effort are held down to a snail's pace by this mass of uninterested unacademic students. The course is thus a compromise and satisfies neither group.

How much better for all concerned if there are ample courses from which to choose and a wise counselor to advise the students! The results of wise counseling are clearly apparent in the accompanying table.¹ The guided and the unguided groups were of about the same average I.Q., 105 and 108. The effect of guidance is reflected in the number of subject failures and the number out at work.

| | Out at work | Out by transfer | Failed one subject | Failed two or more subjects |
|---------------|-------------|-----------------|--------------------|-----------------------------|
| Guided..... | 4.5 | 9.1 | 18.2 | 0.0 |
| Unguided..... | 12.1 | 13.1 | 30.8 | 10.3 |

¹ Proctor, W. M., *Psychological Tests and Guidance of High School Pupils*, Journal of Educational Research Monographs, No. 1. Bloomington, Ill.: Public School Publishing Company, 1923.

The reduction of failures for one subject from 30.8 per cent to 18.2 per cent and for two subjects from 10.3 per cent to 0 per cent is especially noteworthy. The bases of guidance in this study were far broader than intelligence-test scores, but these latter undoubtedly entered into the advice concerning the choice of subjects.

Intelligence tests are also used to help students decide on various courses of study. Since the members of some courses have a much higher average intelligence than others, this information should be conveyed to the student who is soon to enter them. In one case,¹ the average I.Q. for the general course was 114.5; commercial course, 109.4; technical course, 108.9; industrial arts course, 103.1; dressmaking course, 97.4. These I.Q.s represent very well the relations between the courses selected and the I.Q. scores. A corresponding report from the city of Saint Louis gave to those in the scientific course an average I.Q. of 109.8; in the general course, 106.3; classical, 106; commercial, 103.2; manual training, 102.5; art, 102.1; and home economics, 100.5. These two studies make clear that the scientific, language, college-preparatory courses enroll on the average more intelligent students than the other courses. This same trend is noticeable at the college level. At Ohio State University and at the University of Illinois the arts, commerce, and journalism drew from those better equipped in intelligence than veterinary medicine, dentistry, and pharmacy.² The average scores on the intelligence tests were also high in medicine, law, and engineering. These divisions of the university were closely alike, varying only from 141 to 147. Veterinary medicine with a score of 112, dentistry with 115, and pharmacy with 125 are poorest of all in their intellectual requirements. A student who would rank in the lowest quarter as a student of law might be above the average of his classmates in dentistry. He would then have to decide whether to struggle along in law or to shine in dentistry.

USE OF INTELLIGENCE TESTS IN HOMOGENEOUS GROUPING

For many years teachers have realized the difficulties inherent in attempting to teach pupils or students who are widely different in their capacities to learn. Those explanations and materials which were suitable for the average of the class would bore the bright and confuse the dull. If the teacher pitched her class discussions and materials on the level of the bright most of the class would be doubly confounded. To remedy the situation homogeneous grouping of pupils or students has

¹ Clark, R. S., "Some Results of Psychology Tests Given to Groups of Public School Pupils of N.Y.C." *Contributions to Education*, Vol. 1, pp. 98-116. Society for Experimental Study of Education.

² Pintner, Rudolph, *Intelligence Testing*, 297. New York: Henry Holt and Company, Inc., 1931.

been suggested and tried out in many schools. In this procedure an intelligence test is given to a large number of students; then, as is ordinarily the case, three sections or classes are made: (1) those who score in about the upper 20 per cent, (2) the 60 per cent falling next, and (3) the lowest 20 per cent. On the very face of it the homogeneity of any group is not marked. The upper 20 per cent may include those from I.Q.s of 112, a bright pupil, to 140 and above, most certainly a gifted child. It must be realized also that a score on an intelligence test is an average of seven or eight subtests, such as arithmetic reasoning, opposite of words, sentence completion, and picking out the best reason. Thus the same median score might be obtained by one pupil who was good in arithmetic reasoning and poor in language and by another whose case was the opposite. These arguments make it clear that these students, made homogeneous on the basis of their intelligence-test scores, are really not as much alike as they would seem. On the other hand, there is undoubtedly greater homogeneity than would obtain in the three groups combined. Humanitarians claim that to label a slow group by calling them "the Z group" or the "opportunity class" is decidedly undemocratic. These persons also believe that this procedure of segregation causes the slow to be more conscious of their plight and hence may develop in them a feeling of inferiority. Claims also are made that since life has in it dull, normal, and superior individuals, all of whom must learn to get along together, the school also should group them that way.

Those who favor homogeneous grouping are activated by the following facts. Those of superior intelligence stimulate each other much more and are apt to accomplish more work if they have as competitors those of the same intellectual level. Their progress depends upon the ingenuity of their instructor in providing a more advanced type of material and in demanding deeper and broader understandings of topics which they investigate. A summary of the many experiments which are concerned with homogeneous grouping gives conflicting results. In general the backward or slow learners profit most by grouping them homogeneously. The rate of progress may be adjusted to their speed of learning, and explanations can be illustrated with more concrete details. The middle group are not much affected, and the success of the superior group depends upon whether the teacher is willing to change the materials and procedures to those more suited to superior intellectual levels.

AIDS IN MAKING DECISIONS ABOUT GOING TO COLLEGE

Intelligence tests furnish evidence bearing on the subsequent success of a high school student in college. Like many another factor which helps to constitute the total prediction picture, its scores only point in certain directions. The coefficients of correlations have been computed perhaps

a thousand times or more between college marks and intelligence-test scores. In the vast majority of cases they have ranged from .35 to .60. Under normal conditions one can confidently expect a coefficient from .4 to .6 between test scores and the average of school marks.¹ Coefficients as high as .6 reduce the error of estimate by 20 per cent. If we made a prophecy based on such a correlation our prophecy would be roughly 20 per cent better than if we had not used the test. However, the test is much more efficient than this. Suppose a student had consistently worked hard in high school but still had made only fair grades. Suppose that his I.Q., based on an intelligence-test score, was only 90. This corroborative evidence might be the deciding factor, for certainly a person who had done his best in high school and even then was only able to pass would have rough going in college. If on the other hand, a student, who has frittered away his time in high school and passed, but was shown by the test to have an I.Q. of 110, would be much more likely to succeed in college did he suddenly acquire a new motive.

Better predictions of subsequent college success can be made by combining several factors than by the use of any one of them singly. In a bulletin from the University of Wisconsin² correlations are published between grade-point averages and such intelligence tests as the Ohio State Psychological Examination and the American Council on Education Psychological Examination. The coefficients computed with large numbers of subjects ranged from .41 to .61. By combining intelligence-test scores and marks for the senior year in high school, a multiple coefficient of .71 was secured. This raises the predictive efficiency from 20 per cent, secured from a coefficient of .60, to 30 per cent, secured from a coefficient of .71.

Intelligence tests have been used to *define more accurately the levels of feeble-mindedness*. Before the advent of tests, feeble-mindedness was defined in terms of the prudence with which one managed his ordinary affairs, the skill with which he adjusted himself to his environment, or his capacity to make a living. While these concepts of feeble-mindedness are still influential in some quarters, definition in terms of M.A. or I.Q. secured from a standard intelligence test is gaining ground. Using the test as the criterion of judgment, we may define the level of feeble-mindedness as shown in the accompanying table. There is pretty general

¹ One factor which keeps these correlations low is the unreliability of school marks. Such unreliability is reflected most dramatically in the wide variations of the coefficients in single subjects. When the marks for all subjects are combined into some such unit as the point-hour ratio, the standing as determined by school marks becomes very reliable.

² Froehlich, Gustav J., *The Prediction of Academic Success at the University of Wisconsin*. Madison: Bureau of Guidance Records, University of Wisconsin, 1941.

| Level | M.A. | I.Q. |
|---------------|-------|-------------|
| Idiot..... | 0-2 | 0-20 |
| Imbecile..... | 3-6 | 20-40 |
| Moron..... | 7-8-6 | 40-65 or 70 |

agreement about these definitions of idiot and imbecile and about the beginning of the limits of the moron. Disagreement arises on the upper level of the moron. Professor Pintner recommended that the upper limit be placed at 8 years and 6 months and that of the upper I.Q. at .60. The upper limits of the I.Q. are dependent upon the C.A. which shall be used in the denominator in cases of maturity. As has been indicated (pages 362, 363), 14, 15, and 16 have been used as ages of maturity. This means that the determination of the I.Q. of a boy at 15 and above would use either 14, 15, or 16 as chronological age. Terman uses an I.Q. of 70 as the dividing line between normality and feeble mindedness. Wechsler uses an I.Q. of 65 for this same purpose. According to him the limits of a moron would extend from 7 to 10-6, with I.Q. limits of 40 and 70. In the Terman-Merrill Revision, beginning at year 13, 1 month is subtracted from each 4 months of C.A. until they reach 15. If we use 10-6 as the upper limit, many more of the population would be included in the feeble-minded category than when 8-6 is used. Pintner reports:¹

Similarly when applying these limits to a random sampling of 4,925 school children not including children in special classes I find that 1.3 per cent fall below I.Q. 60 and 6.6 per cent below I.Q. 70. It is, therefore, probably wiser to consider the upper limit of feeble-mindedness as lying somewhere in the neighborhood of I.Q. 60 and M.A. 8-6.

The present author agrees with this recommendation.

USES OF INTELLIGENCE TESTS FOR VOCATIONAL GUIDANCE

The leading problems here are (1) the discovery of the amount of intelligence required for successful competency in any given occupation, (2) the measuring of the intelligence possessed by the individual in question, and (3) the guidance of the individual into the vocation for which his intelligence fits him. In solving the first problem it would seem a simple procedure to get an unselected sample of machinists, for example, to take two or three intelligence tests, and then compute their median and percentiles. In such a manner the intellectual requirements of occupations could be determined. Only in this way could

¹ Pintner, *op. cit.*, pp. 340-341.

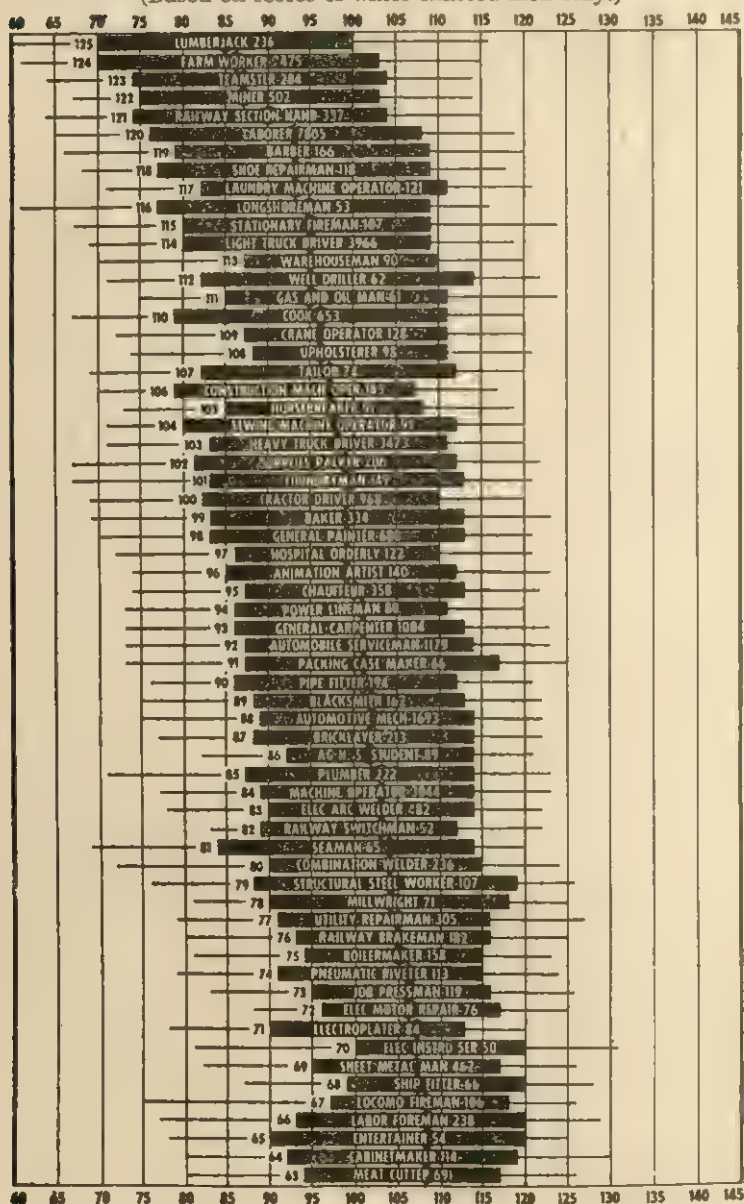
adequate standards be secured. Instead of this direct procedure the data have had to be gathered indirectly from tests administered to draftees in two world wars. Such medians and percentiles as we have accumulated are computed from the records of those who *said* they were machinists—a procedure subject to error, because an individual who was merely a machine tender sometimes puts down his occupation as machinist. In some occupations, such as clerical workers, engineers, lawyers, and doctors, intellectual requirements have been clearly defined in terms of *intellectual units*. In the vast majority of occupations, however, these defined requirements in intelligence have not been determined.

A second difficulty arises from the nature of the intellectual requirements in various vocations. Wherever large numbers of subjects in a given vocation have been tested, wide variations in intelligence have been found. A part of this difficulty has arisen because not enough concern was given to the degrees of competence reached within the occupation. Suppose that “machinists” was the classification in question; then a part of the variation in intelligence could be attributed to the fact that some of the members of this occupation were apprentices, some journeymen, and still others experts. It is also evident that weakness in intelligence among the workers in a given occupation can often be overcome by industry, a pleasant smile, and tact. In many cases a person of intelligence inadequate for real success in the occupation drags along at the lower end of the procession. Striking illustrations of incompetency appear even in such highly organized professions as law and medicine.

These variations in the intelligence required for successful competency in occupations cause so much overlapping in test scores that the 25th percentile in that occupation with the highest intelligence requirements will frequently fall at the 75th percentile of one far below it. For example, as based on data from the First World War, the 75th percentile of the electrician was 109 points on Army Alpha, while the 25th percentile of the physician was 107. This means, of course, that the upper 25 per cent of electricians were on the Army Alpha as good as, or better than, the 25 per cent of physicians just below the average. Many electricians no doubt had intelligence scores above the average of the physicians. This factor of overlapping of intelligence required in various occupations makes the guidance problems much more involved.

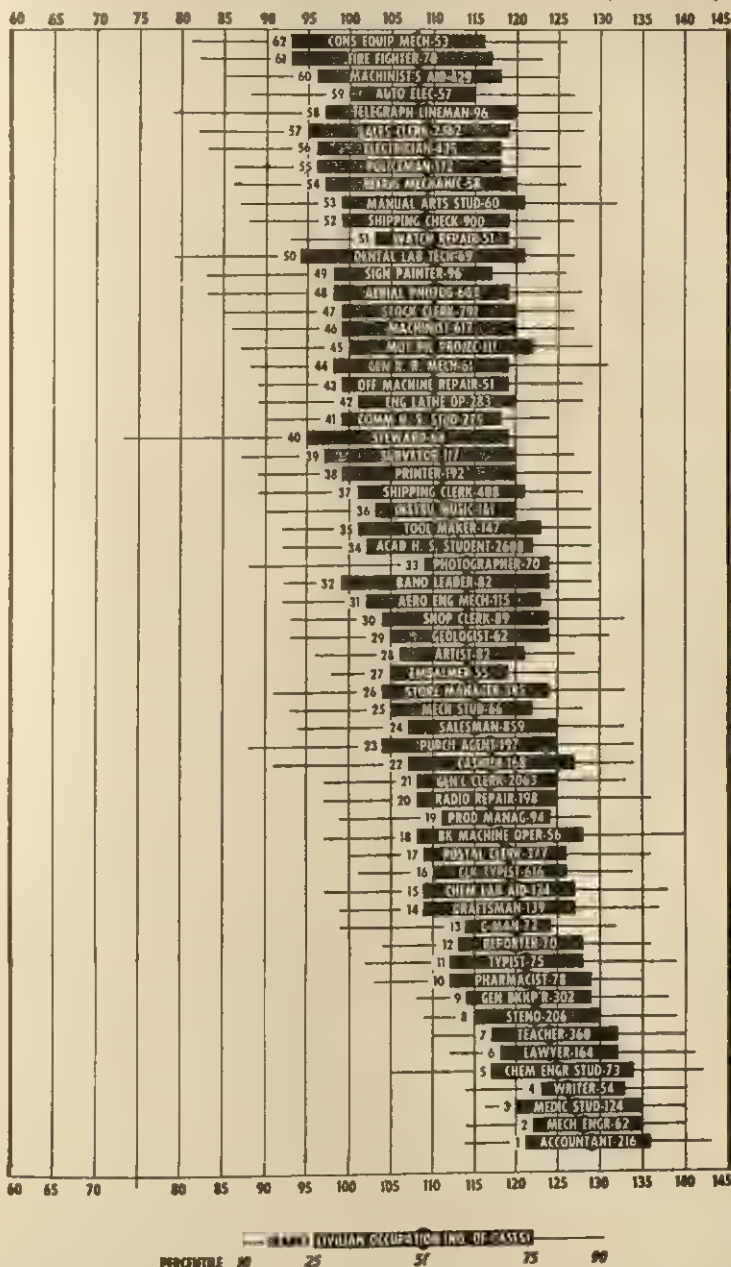
It was suggested in the previous paragraphs that the intelligence scores for occupations as listed in the army draft of 1917 must be accepted with some reservation. The army draft was interested in disrupting essential industries as little as possible. Let us look at those individuals who classed themselves as farmers. Farming was essential for carrying on the war; hence all owners were not drafted. Most of those

TABLE 15. AGCT SCORES FOR CIVILIAN OCCUPATIONS
(Based on scores of white enlisted men only.)



LEGEND Each bar with extensions shows the place of an occupation in the AGCT score range of 60-145 (10). As indicated below, each bar shows the middle fifty per cent of scores in that occupation, and the extensions show the 10th and 90th percentiles:

TABLE 15. AGCT SCORES FOR CIVILIAN OCCUPATIONS (Continued)



Examiner's Manual, First Civilian Edition, November, 1948, page 8. (By permission of Science Research Associates, Chicago.)

classifying themselves as farmers were either hands on the farms or renters. Therefore the intellectual level of the farmers tested was affected by the manner in which the draft worked. If it had worked alike in the case of all occupations the results would not have been so badly affected, but there was a differential effect upon various occupations. The computations of the average, Q_1 (25th percentile) and Q_3 (75th percentile) for each occupation were made by the army psychologists.¹ Additional data have since been accumulated and the army results modified so as to present more accurate intelligence scores for all occupations. Two inferences seem warranted from the army data: (1) there is a hierarchy of intelligence scores in the various occupations, and (2) there is a tremendous variation among intelligence scores possessed by members of any occupation.

Data gathered from the Second World War, while differing in detail, lead to the same conclusions. The Army General Classification Test (AGCT) was developed to classify inductees according to "their ability to learn quickly the duties of a soldier." It is composed of three test forms: (1) vocabulary, (2) arithmetic word problems, and (3) block counting. There are 150 items in all. It does not differ greatly from the usual intelligence test. All raw scores were converted into AGCT standard scores with a mean of 100 and a standard deviation of 20.

Table 15 shows the scores received by men in various occupations on this test. The black bars indicate the range of scores from the 25th percentile to the 75th percentile. The lighter extended lines indicate the range below the 25th and above the 75th percentile. The first thing that strikes the eye is the large differences between the occupations whose members score high and those whose members score low. Consider the teamster and the miner on the one hand and the accountant and mechanical engineer on the other. Observe next (1) the range between the 25th and the 75th percentiles, and (2) the extent of the scores below and beyond these points. Compare the middle 50 per cent of the *seamen* (81 in rank), which is about 30 points, with that of the *writer*, about 10 points. In the third place, observe the overlap between occupations. Some cabinetmakers (rank 64) score as high as 130, which is above the median of the accountants, who rank first. Even some lumberjacks, who rank lowest in this occupational hierarchy, score above 115, which is higher than 11 per cent of the accountants' scores.

From such considerations the suggestions about entering certain occupations obtained from intelligence-test scores must be highly tentative. Suppose a student receiving 120 points on the AGCT wished to enter medicine. You might say to him, "Your chances of success in medicine are rather slim. You rank at the 25th percentile of medical

¹ *Memoirs of the National Academy of Sciences*, Vol. XV, Part III, Chap. 15, 1921.

students. On the other hand your score is at the 55th percentile for pharmacists and the 60th percentile for salesmen."

The usefulness of the tests for guidance varies with the occupation. Success in some occupations is nicely correlated with scores on intelligence tests. Executives' success is closely dependent upon their intelligence.

An intelligence test was given to minor executives in 1915, and again in 1920, and the results compared with the firm rank. The correlation was .69. A small group of executives at the head of a concern were ranked by the vice-president as to their executive ability. The correlation with their rank in an intelligence test was .89.¹

In many types of occupational activity there is almost no relation between intelligence-test scores and success.

SUMMARY

The apparent disadvantages of group tests in comparison with individual tests have been largely overcome. The best standardized group intelligence tests approach very closely the accuracy of the individual test. Out of the needs of the First World War for a rough intellectual classification of a large number of men there developed the Army Alpha. This test for literates, as well as the Beta test for illiterates and those inept in English, were applied in a large number of situations during this war. The data thus collected furnished living evidence of the value of group tests of intelligence. From this beginning the construction of group tests of intelligence went forward by leaps and bounds until today there are carefully standardized group tests available for every age from 5 years to maturity.

The detailed analysis of three series of intelligence tests displayed three types of test construction. In one of them, the Pintner series, careful statistical analysis was made at every stage of the test's construction and development. Just how good a test it is, then, can be easily determined. The Kuhlmann-Anderson group tests were also carefully constructed but leaned more heavily on the subjective judgment of the authors, who had worked for years in tests, than on more refined statistical analysis. The third type, PMA, divides intelligence into five abilities which are fairly independent and computes the reliability of each.

A study was then made of the test forms which had been found useful in constructing intelligence tests suitable for the various grade levels

¹ Burt, Harold E., *Principles of Employment Psychology*, p. 279. Boston: Houghton Mifflin Company, 1926.

(kindergarten, grades 1 to 3, grades 4 to 8 and grades 9 to 12) and samples of items found suitable were introduced. It was disclosed that certain test forms were useful at all stages even though with younger children the relations were expressed in pictures. Analogies, opposites, number completion, logical selection, classification, vocabulary, best answer, and arithmetic reasoning occur again and again. Increasing difficulty is attained by using more subtle and more unusual relations and by making the possible answers more nearly alike.

Intelligence tests, both group and individual, have wide fields of usefulness. In all cases their function is supplementary and only one of a constellation of factors concerning the individual. They do aid in guiding children and students into schoolwork where they can work happily. In this manner, failure is reduced, students like school better, and they stay in school longer. Intelligence tests are useful in diagnosing individual difficulties which aid us in understanding the failure of a student and in planning for him a more successful course of action. Intelligence tests also have proved their worth in defining more accurately the lower and upper levels of intelligence. Feeble-mindedness is now clearly defined in terms of mental age and I.Q. Finally, some help is furnished by intelligence tests to the counselor in the area of vocational guidance. However, the upper and lower limits of intelligence for each vocation have not been determined. This lack limits the usefulness of intelligence tests in this very promising area of guidance.

QUESTIONS AND EXERCISES

1. What criticisms were leveled against the group test of intelligence? How were most of these criticisms met?

2. Describe the salient characteristics of the Army Alpha Test. What test forms were used in its construction?

3. Illustrate the influence of range of subjects on reliability by reference to the Pintner-Cunningham test. Explain the new technique, introduced by Pintner, for computing the I.Q. Compare with the older method.

4. What criticisms were made of the construction of the Kuhlmann-Anderson test? What criteria of validity did these authors use? What weakness appears in their treatment of reliability?

5. How do group tests of intelligence suitable for the kindergarten and entering first grade differ from those intended for grades 5 or 6?

6. Name and illustrate six test forms used in constructing group tests of intelligence for the elementary grades, the upper grades, and high school. What are the common characteristics of all these tests?

7. How are intelligence tests used in the first grade? With children in the primary grades?

8. Discuss the uses of intelligence tests in aiding students to select courses of study. What subjects in elementary school are highly correlated with intelligence-test scores? What subjects have only low correlations with these same scores?

9. What has been the relation between intelligence-test scores and success in school? Continuation in school? Number of subjects failed? How is this problem being met at present?

10. How have group tests of intelligence been used to form homogeneous groups? Give two reasons why these groups are not as homogeneous as they seem. Do you favor such groups? Why?

11. What types of scores furnish the highest prediction of college success?

12. How have intelligence tests been used to define feeble-mindedness? What

are some difficulties present in trying to decide the upper limits of feeble-mindedness?

13. Describe the application of group test scores to the problems involved in vocational guidance. What difficulties present themselves when we attempt to decide the amount of intelligence needed for good performance in any occupation?

BIBLIOGRAPHY

Books

BUROS, OSCAR K. (ed.): *The Third Mental Measurements Yearbook*. New Brunswick, N.J.: Rutgers University Press, 1949.

BURT, HAROLD E.: *Principles of Employment Psychology*, rev. ed. Boston: Houghton Mifflin Company, 1942.

CRONBACH, LEE J.: *Essentials of Psychological Testing*, Chap. 8. New York: Harper & Brothers, 1949.

FREEMAN, F. N.: *Mental Tests*, rev. ed. Chaps. V, VI. Boston: Houghton Mifflin Company, 1939.

FROELICH, GUSTAV J.: *The Prediction of Academic Success at the University of Wisconsin*. Madison: Bureau of Guidance Records, University of Wisconsin, 1941.

JORDAN, A. M.: *Educational Psychology*, 3d ed., Chap. 13. New York: Henry Holt and Company, Inc., 1942.

KOOS, LEONARD M., and GRAYSON N. KEFAUVER: *Guidance in Secondary Schools*. New York: The Macmillan Company, 1932.

Memoirs of the National Academy of Sciences, Vol. XV, Part III, Chap. 15, 1921.

MONROE, W. S. (ed.): *Encyclopedia of Educational Research*. New York: The Macmillan Company, 1941.

PINTNER, RUDOLPH: *Intelligence Testing*, Chaps. VII, VIII, XII. New York: Henry Holt and Company, Inc., 1931.

PROCTOR, W. M.: *Educational and Vocational Guidance*. Boston: Houghton Mifflin Company, 1925.

Articles in Journals—Manuals

DAVIS, H.: "Intelligence Tests in Public Schools in Jackson," *Twenty-first Yearbook of the National Society for the Study of Education*, Chap. III, pp. 131-142. Bloomington, Ill.: Public School Publishing Company, 1922.

DURRELL, DONALD D.: "The Influence of Reading Ability in Group Intelligence Measures," *Journal of Educational Psychology* (1933) 24:412-416.

FRYER, DOUGLAS: "Occupational Intelligence Standards," *School and Society* (1920) 16:275.

JORDAN, A. M.: "Student Mortality," *School and Society* (1925) 22:821-824.

—: "The Validation of Intelligence Tests," *Journal of Educational Psychology* (1923) 14:348-366, 414-428.

Kuhlmann-Anderson Tests, Instruction Manual. Minneapolis: Educational Test Bureau, 1944.

MADSEN, I. N.: "The Contribution of Intelligence Tests to Educational Guidance in High School," *School Review* (1922) 30:672-701.

PINTNER, RUDOLPH: *Manual for Administering and Scoring the Intermediate and Advanced Tests*. Yonkers, N.Y.: World Book Company, 1943.

POWERS, S. R.: "Intelligence as a Factor in the Election of High School Subjects," *School Review* (1922) 30:452-455.

STEWART, NAOMI: "AGCT Scores of Army Personnel Grouped by Occupation," *Occupations* (1947) 26:5-41.

PART THREE

Personality Inventories

CHAPTER 16

Measurement of Interest

It is important to know what activities either in reality or in imagination have left an individual with a glow of satisfaction, *i.e.*, which ones he has found interesting. The importance of this knowledge arises out of the fact that real happiness in life comes from doing well what is enjoyed, and out of the fact that if an activity arouses interest it will be pursued with less friction and with more likelihood of success. It is important also because in exploring various areas of interest one may sometimes discover new interests which before were not realized.

Moreover, the discovery of areas of interests in school children may not only furnish teachers with information so useful in motivating and selecting children's curriculums within the class but also may aid the counselor in assisting his clients to come to some decision concerning the courses they will take in school and the occupations they will enter. Here is not the place to discuss the relation between interest and learning, but these two are inextricably intertwined.

Except in a very broad way no one has set down a list of desirable interests which a student should possess. It is, therefore, impossible to measure the degree of success in those interests which are the objectives of teaching. The problem in the measurement of interests becomes, then, one of discovery for purposes not of evaluation but of guidance.

CHARACTERISTICS OF INTERESTS

Interest and motive are closely related. The interest which an individual has in an object frequently arouses the motive of acquiring it. Motive, as defined by Professor Woodworth, is a "state or set of the individual which disposes him for certain behavior and for seeking certain goals."

Note that a motive is not the situation or the stimulus, but a *set towards a certain goal*. Thus, a motive releases energy and directs it. Hunger is the motive. The food is the incentive which releases a larger or smaller amount of energy in accordance with its attractiveness. It is the attractiveness of the goal to the individual which arouses the motive and which gives one incentive prepotency over

another. John Dewey brings the set and the motive together in his description of the latter as a "wholehearted identification of oneself with a goal or activity."

Tied up closely with motive is the matter of interest. *Interest is the pleasant feeling tone which attaches itself either to the activity or to the goal.* If it attaches itself primarily to the activity, it may be called *intrinsic*; if primarily to the goal, *extrinsic*. *Along with this feeling tone there is also in interest an urge to continue the activity or to seek the goal.* Intrinsic interest arises either because the activity connects up directly with these inherited body needs such as hunger, sex, thirst, fear, anger, bodily activity; or, because it falls in directly with habit-patterns already started. Extrinsic interest arises because of anticipated satisfaction in the goal itself.¹

The major type of interest with which we are presently engaged is the intrinsic interest. Intrinsic interest leads to a return to the experience and a dwelling upon it. It accounts for those peculiar anomalies in which a person reaches in some undertaking to higher levels of attainment than might be expected from his moderate capacity or, conversely, the lack of which causes an individual of great capacity to attain to only mediocre success because his heart (interest) is not in his work. Later on in this chapter we shall offer evidence of the relation between achievement and interest, but at the moment it is sufficient to say that capacity and interest are mutually supplementary. It is the presence of both which brings the highest success. If we can discover those lines of activity which fill full, continue, or extend the ongoing activities of individuals, their success may be greater than expected.

METHODS OF DISCOVERING INTERESTS

The most direct method of discovering interests, of course, is *to go directly to the subject* and ask him what he likes, what he is interested in. Here we assume that his cooperation is already attained before the start is made. The complexity of this procedure may vary all the way from a simple question, such as "List three books which you like very much," to a set of three or four hundred questions drawn from a rich variety of human activity. All these methods are limited in several ways. In the first place *truthful answers depend upon the willingness of the subject to cooperate*. If the answering of the question compromises the subject in any particular he is not apt to answer truthfully. Thus the authors found that, in listing the names of five most interesting books, 5,000 high school students rarely listed the names of salacious books or

¹ From Jordan, A. M., *Educational Psychology*, 3d ed., pp. 154-155. New York: Henry Holt and Company, Inc., 1942. By permission.

magazines, which unfortunately are frequently read. Spencer¹ found that unsigned questionnaires were answered with more openness and frankness than they would have been had they been signed. The second difficulty, especially applicable to the elaborate questionnaire about many occupations, is simply *a lack of information* about the vocation or activity. How could a student really choose which activity he likes best and which worse from these questions?²

g. write novels

h. conduct research on the psychology of music

i. make pottery

He knows nothing of writing novels or of making pottery, and as for what "research on the psychology of music" is, he has not the slightest idea.

Even worse possibly than total ignorance is the generalization about occupations which is frequently made from a few glaring instances. A subject is presented with a list of occupations about which he is to express his interest. His imagination has been fired by the extreme incomes which certain workers in those occupations have made. But the great lawyer's \$50,000 fee for one case, and the baseball player who receives a salary of \$70,000 a year are glittering illustrations of what the incomes in these occupations usually are not. The student too much influenced by exceptional cases fails to consider what income the average and lower bracket of workers receives in these occupations.

Another method which at first glance seems very promising is that of *direct observation*. This method has been tried out in a number of situations. The author, for example, used observation to check the interests of children in books and magazines against their interests as expressed through the questionnaire. He found, for example, that children wore out the copies of some books in libraries while others were clean and fresh, that certain cards in the card catalogue were dog-eared and dirty, that certain books were hidden behind radiators and the poetry shelves so that they could be secured on the children's next visit, and finally, that certain books were chosen first by a large number of children day after day. In another case, anecdotal records of activities of an individual, such as those involved in helping solicit money for the school annual or in fashioning a lampstand, have been recorded and used. In some schools, short introductory courses embodying some of the essential features of occupations have been tried out and records made of the

¹ Spencer, Douglas, *Fulcrum of Conflict*, p. 192. Yonkers, N.Y.: World Book Company, 1939.

² Kuder, G. Frederic, *Preference Record*, Chicago: Science Research Associates, 1942.

apparent pleasure with which these activities were undertaken and finished. This is a promising if comparatively undeveloped field of interest discovery.

The third method is based on the assumption that the *greater the amount of information* which an individual possesses in any area the *greater will be his interest*. The idea here is that if the student likes a certain area he will read more about it, work at it longer, and remember it better than he does in those areas where no interest is present. Here again the opportunity for acquiring information rather than the interest might have been lacking.

The most successful procedure for discovering the interests of students and adults has been that of direct questioning. Despite the many possibilities of errors inherent in the method, this technique, based on asking the subject directly whether or not he has any interest in a presented item, has proved most successful. Four inventories are first presented in some detail in this chapter followed by a selected list of inventories which have been used to discover interests. The Strong Vocational Interest Blank will first be presented. It is soundly constructed and has been carefully revised.

INTEREST INVENTORIES—QUESTIONNAIRES

Strong's Vocational Interest Blank is the oldest of these measures of subjective inventoried interests. Its origins go back to the work of a seminar in 1919 at the Carnegie Institute of Technology. At this seminar under the direction of Clarence S. Yoakum, a group of graduate students and professors began to gather interesting items which distinguished between members of different occupations. There were subsequent attempts to organize these items into usable inventories for purposes of (1) distinguishing between the interests of bright and average children, (2) distinguishing between various social groups by their interests, and (3) distinguishing between engineers whose work was (a) mechanical in nature and, (b) social in nature. As these inventories were developed improvements were made in scoring. The number of degrees of interest or liking varied from yes—no—0 (do not care) through L—1—?—d—D (like very much, like, not decided, dislike, dislike very much) to a scale with seven divisions: (1) very strong dislike, (2) marked dislike, (3) some dislike, (4) indifference, (5) some liking, (6) marked liking (7) very strong liking. Strong's Vocational Interest Blank uses three divisions L (like), I (indifferent), and D (dislike) commonly called the L-I-D procedure. Furthermore, Strong not only launched his blank or inventory but studied it, revised it, and improved it. The principle of construction of this instrument is based on the tendency of men to gravitate to the occupation which they like and as

a result to have certain common interests which can be tapped. The procedure of construction consists of selecting a large number of items which will differentiate between "men in general" and men successful in a certain occupation. Items which do not distinguish between these two groups are thrown out. These items are then weighted in scoring in proportion to the degree of completeness with which they separate these two groups.

Let us illustrate from Strong's blank. In computing the score on the item "actor" for personnel managers the following procedure was used:

| | Per Cent | | |
|--|----------|-----|-----|
| | L | I | D |
| Personnel managers..... | +49 | +38 | -13 |
| All others..... | +38 | +35 | -27 |
| Difference..... | +11 | +3 | -14 |
| Final weights for item of "actor"..... | +2 | +2 | -3 |

Strong worked out a scheme whereby a difference of 8 to 11 was given a weight of 2, one of 3 to 7 a weight of 1, and one of 12 to 15 a weight of 3.

Form M, as at present constituting the Strong Vocational Interest Blank for Men, is composed of 400 items divided into the following parts:

| Part | Number of Items |
|--|-----------------|
| I. Occupations..... | 100 |
| II. School subjects..... | 36 |
| III. Amusements..... | 49 |
| IV. Activities..... | 48 |
| V. Peculiarities of people..... | 47 |
| VI. Order of preference of activities..... | 40 |
| VII. Comparison of interest between two items..... | 40 |
| VIII. Rating of present abilities and characteristics..... | 40 |
| Total..... | 400 |

It is now possible to score these 400 items differently for each of 39 separate occupations. For each occupation there is an appropriate scoring key with the items scored differently for each occupation. The number of persons included in these samples for each occupation ranged from 113 representatives of the occupation of YMCA secretaries to 513 engineers. In 13 out of the 35 occupations the number in the sample was 250 or more. These occupations vary from architect to lawyer to real-estate salesman and from chemist to mathematician to physicist. There are procedures also provided for scoring certain groups of occupations so that all 39 would not have to be scored. Scales have also been developed for securing measures of (1) maturity of interests, (2) occupational

level, (3) self-confidence and sociability, (4) social adjustment, (5) scholastic aptitude, and (6) theoretic and economic evaluative attitudes.

The reliability, validity, norms, and manual have all been carefully worked out. The reliability, as indicated by the mean coefficient of 21 occupations, is in the neighborhood of .877 as computed by the odd-even technique when 285 Stanford seniors constituted the population. With this same group of seniors the coefficient was .75 after a period of 5 years. Using the test-retest method the coefficient after the lapse of 1 week was .869. With high school students the reliability is somewhat lower. One study (Carter, Canning, and Taylor, 1941) states, "Thus if a high school boy receives a 'C' rating on the first test, there is an 83 per cent chance that he will receive the same rating and only a 1 per cent chance that he will receive an 'A' rating two years later." Then, too, if he receives an "A" rating in any occupation on the first test there is an 88 per cent chance that he will receive an "A" or "B" rating two years later. This reliability of .87 is to be compared with a coefficient of .94 or .95 on our best intelligence tests and .95 to .97 on our best educational tests. One must remember that the reliability efficiency based on a coefficient of .87 is about 51 per cent, while one based on a coefficient of .95 is 69 per cent efficient or dependable. One can conclude that this test is fairly reliable for the diagnosis of a single individual's interest.

The validity of a test is usually difficult to determine and especially so in the case of interest inventories. Strong argues that his inventory is valid because of the manner in which it was constructed. In selecting the men from whom to differentiate the interests of "men in general," the greatest care was taken to make certain that they really represented the occupation in question. For example, only the successful members of an occupation who had worked in that occupation at least three years were used. The average age of the individuals in these occupational groups was 43 years. In the second place, Strong argues for the validity of his inventory along three lines. In the first place, among the 933 nonengineering men at Stanford, only 15 per cent rated an A on the engineering interest scale, while of those taking engineering 75 per cent rated A. In the second place, there is considerable relation between the amount of interest as indicated by the blank and success in some occupations. For example, among life-insurance agents 67 per cent of those who scored A sold at least \$150,000 of insurance in one year, while only 6 per cent of those receiving C did so. In the third place, those who continued in an occupation scored higher interest ratings than those who dropped out of that occupation. In general, when men changed from one occupation to another they tended to score higher in the second occupation. Such lines of evidence, together with the fact

that over the years the clinical use of the blank has borne out this contention, convince one of the validity of this inventory.

Norms of the inventory have been worked out in letter scores, standard scores, and percentiles. Sample tables of distribution are furnished in the manual. The most practical one of these measures is the letter score. If a subject's interests agree pretty largely with those of a certain occupation he receives an A in that occupation. Technically, if a subject's interest score is not lower than 0.5 sigma below the average of an occupation then he receives an A. This amounts to 69 per cent of the highest scores in that occupation. If he falls in the next 29 per cent of that occupation's score, he receives a B or B—. In the lowest 2 per cent he receives a C. An individual who scores a C in any occupation has no real interest in that occupation, or no more than "men in general." The method of scoring, reliability, validity, and norms are all clearly explained in the manual.

Strong has also issued a Vocational Interest Blank for Women built in the same way as the blank for men. It contains 400 items, 263 of which are the same as those contained in the blank for men. There were in 1951 19 occupational scales ready for use varying from "artist" and "author" to "teaching physical education in high school" and "YWCA General Secretary." Reliabilities, validities, and norms are developed in a manner similar to those for the men.

One of the great difficulties with this inventory is the time it takes to score. If scored by hand, even by an expert, it takes 5 to 10 hours to score the 39 different occupations. It is also expensive to have the blanks scored by machines at the central office. Some experimenters have tried to simplify the scoring by weighting the answers 1, 0, -1 in all cases instead of using the present scheme in which the score for an L-score ranges from 4 through zero to -4 and for a D-score from +4 to -4. Strong holds that this procedure makes the results a little less reliable and hence will have none of it.¹ In the second place, a liking or disliking of an item may be superficially acquired. It may be based on one experience with it and hence the generalization may be specious, or it may be due to a lack of information about the activities in question, or there may even be an attempt on the part of the subject to prevaricate about his real interest. For these reasons no one should fill out the blank who is not seriously concerned about arriving at a knowledge of what his real interests are.

¹ Strong, Edward K., Jr., "Weighted vs Unit Scales," *Journal of Educational Psychology* (1945) 36:193-216. Strong says (p. 215), "On such a basis unit scale scores will lead to different counseling from weighted scores in from one-sixth to one-twelfth of the cases."

The Cleeton Vocational Interest Inventory approaches the problem of interest in a manner similar to that of Strong's Vocational Interest Blank. It lists occupations, school subjects, characteristics of people, activities, and magazines and asks the subject to express his likes or affirmations by placing a + after the item and his dislike or negation by placing a 0 in the same position. There are 670 items in all, grouped around nine occupational families and an introvert-extrovert dimension. The areas or families of occupations are (1) physician, (2) life-insurance salesman, (3) engineer, (4) teacher, minister, or social worker, (5) purchasing agent, (6) lawyer, (7) mechanical occupations, (8) accountant, statistician, or banker, (9) actor, musician, or artist. The occupations, activities, school subjects, characteristics of people, and magazines whose liking would be customary in this type of occupation are collected in that occupational type which is being studied. For example, under the heading AA (physician) three groups of items (A,B,C) are placed. Under Group A are listed 20 occupations such as bacteriologist, chemist, drug manufacturer, pharmacist, physician, and surgeon; under Group B are placed 20 such items as anatomy, botany, zoology, pet animals, sick people, nervous people; while under Group C come 20 such activities as (1) working for yourself instead of others, (2) ability to meet emergencies quickly, and (3) being a member of a professional society. The last division of the inventory consists of a set of 40 questions purporting to discover the amount of introversion-extroversion.

Corresponding to the inventory for men there is also one for women made in exactly the same way. Its occupational areas are (1) clerk, stenographer, or typist, (2) retail-store salesclerk, (3) nurse or bacteriologist, (4) social worker, vocational counselor, secretary, or lawyer, (5) artist, writer, designer, composer, (6) grade school teacher, (7) high school or college teacher, (8) manicurist, actress, or dancer, (9) housekeeper, factory worker. The *reliability* of this test is satisfactory. By the odds-even method correlations range from .85 to .91 when 150 to 1,000 cases are used. Furthermore, "On a second administration within a month of the first marking of the inventory, 6.1% of the responses were changed from '0' to '+' or from '+' to '0.'" The author holds properly that if it can be shown that the items of the inventory are selected in such a manner that they have significance for specific occupations then their *validity* is assured. He thus selected some items whose basic occupational significance had already been determined as well as some new items because of their agreement with those basic items.¹ The scores of 7,424 persons "successfully engaged in standard occupations" were analyzed in order to determine their standings on the nine scales of the inventory. The results showed a high agreement

¹ *Manual*, pp. 20-21.

between the occupation being followed and the corresponding scale score:

Among these 7424 persons, the highest inventory rating of each agrees with the occupation being followed in 76% of the cases. Eighty-two per cent rate either first or second on the inventory scale corresponding to their occupation, and 95% rate first, second, or third in the corresponding scale.¹

Norms for grades 9, 10, 11, and 12, college freshmen, and adults are available.

This inventory has several strong points. It gives the subject an opportunity to express his likes or dislikes about a very large number (630) and a large variety of items. The inventory is easy to score. One may simply count the number of plusses or it may be machine-scored. Its manual of directions is excellent, describing as it does the development and construction of the scale as well as the conditions under which the test may be used most successfully. Many counselors would agree that the determining of areas of interest is about as far as it is practical to go with an inventory of occupations. But not all features of the inventory are desirable. Many occupations are so much like one another that the subject may carry over his interest in one occupation to the next one listed. Some critics have voiced their objections to the failure of the author to describe his principle of classification whereby only nine areas are arrived at. It is indeed curious to place manicurist with actress and dancer in one category or to place watchmaker under biological sciences. There are no correlations computed between the groups so that one cannot tell whether or not there is overlapping between them. Such a weakness should be rectified. In conclusion, we can say that this inventory is practical and useful and that it furnishes roughly the subject's area of occupational interests.

Kuder Preference Record is suitable from grades 9 to 16, *i.e.*, for both high school and college students. Interest is expressed by indicating a preference among three activities. The statements of these activities are arranged in groups of threes. The instructions are:²

Read over the three activities of each group. Decide which of the three activities you like *most*. Note the letter in front of it and punch a hole through the 1 beside this letter in the column at the right, using the pin with which you are provided. Then decide which activity you like *least* and punch a hole through the 3 beside the corresponding letter in the column at the right.

¹ *Manual*, pp. 21-22.

² Quotation and items by permission of Science Research Associates, Chicago.

The two following triplets will serve as examples:

- | | | |
|---------------------------------------|-------|-----|
| g. Study physics | (1) g | (3) |
| h. Study musical composition | (1) h | (3) |
| i. Study public speaking | (1) i | (3) |
| | | |
| r. Make a study of flower arrangement | (1) r | (3) |
| s. Make a study of mental ills | (1) s | (3) |
| t. Make a study of propaganda methods | (1) t | (3) |

Altogether there are 168 triplets. Underneath the column of answers through which pins are to be punched indicating the most and least liked activity is an ingeniously arranged set of patterns made by small circles connected by a line. If the punched pinholes in the answer sheet fall into the circles constituting a particular pattern, then this pattern receives a score which depends on the number of holes punched. There are nine such patterns: (1) mechanical, (2) computational, (3) scientific, (4) persuasive, (5) artistic, (6) literary, (7) musical, (8) social service, and (9) clerical. It is possible to compare these categories with four types of interest developed from the Strong Vocational Interest Blank. The analysis of the Strong Vocational Interest Blank, in which there is an opportunity for the subject to compare his interests with those of some 39 occupations, showed four outstanding types of interests: (1) science, (2) language, (3) people, and (4) business. We might from the Kuder categories place the scientific area beside science, the literary area beside language, the social-service and persuasive areas beside the interest in people, and the computational area beside interest in business. It is important to note that these two instruments for measuring interest, developed in such different ways, should have come to as much agreement as is here indicated.

This preference record is *reliable*. The reliability of each of the nine divisions has been studied with graduate students, college students, high school seniors, and even with grade 8, and with both men and women, boys and girls. In the vast majority of reliabilities the *r*'s are .90 and above. In only the persuasive at the level of the high school and grade 8 is there inadequate reliability for individual analysis. Here the *r*'s run .82, .80, and .84. The category of mechanical interest is probably the most reliable and that of persuasive interest, the least.

Norms of interest have been established for both boys and girls, for men and women. The present profile sheet was derived from 515 college students composed of both men and women. In the 1944 manual, norms are furnished for sophomore, junior, and senior high school classes for both boys and girls. These norms, based on 500 cases for each age group, are much more satisfactory than the original norms. It seems better to

have separate norms for the two sexes because of substantial sex differences in the mechanical, computational, scientific, musical, artistic, social-service, and clerical divisions. Boys are clearly more interested in the first three groups and girls in the last four. In the literary and persuasive divisions the differences are small. It is quite clear that data are available for comparing an individual's preference record with others of the same level of advancement. Norms are continually being improved by the addition of new cases. Each new manual includes improved bases of comparison.

Considerable resemblance exists between comparable areas of the Kuder Preference Record and the Strong Vocational Interest Blank. For example, Strong's artist score correlates .56 with the artistic area of Kuder (N , 166); Strong's engineer score correlates .72 with the mechanical area and .54 with the scientific area of Kuder; Strong's chemist scores correlate .51 with the mechanical area and .73 with the scientific area of Kuder. With Kuder's computational area Strong's scores of the accountant (C.P.A.), purchasing agent, and banker correlate between .38 and .49, while these same occupational areas of Strong correlate between .36 and .62 with Kuder's clerical interest.¹ While these correlations are substantial it is not possible to interchange their scores. *Their categories are different and must be so considered.*

The newest of these interest inventories suitable for high school students is the Occupational Interest Inventory by Edwin A. Lee and Louis P. Thorpe. The test consists of 120 paired items and 30 items of triads. Two items follow from the 120 pairs in which the directions are: "Put a circle around the letter preceding the activity you choose."²

2

- 19 E Clip hedges and trim trees
- C Mix cement, or carry plaster or bricks

3

- 36 D Check the accuracy of financial statements or records
- F Use scientific laws to develop new machinery

The instructions for the triads are "You are to choose one of the three in each group. Indicate your choice by a circle around the letter preceding the activity." One illustration is:

- 10 a. Keep the accounts and collect the money for a paper route
- b. Manage the financial accounts and collections in a large company
- c. Figure payrolls, salary rates, and salesmen's commissions

¹ The coefficients in this paragraph are from Triggs, Frances Oralind, "A Further Comparison of Interest Measurement by the Kuder Preference Record and the Strong Vocational Interest Blank," *Journal of Educational Research* (1943-1944) 37:538-544.

² Items by permission of California Test Bureau, Los Angeles, Calif.

By scoring the first set of 120 items, interest scores of six occupational families may be obtained:

1. Personal-social (domestic, personal, social services, teaching, law)
2. Natural (farming, gardening, fishing, lumbering, caring for animals)
3. Mechanical
4. Business
5. The arts
6. The sciences

Thus with 120 pairs, or 240 entries, each occupational family has 40 items. Of these items, 10 indicate interest in the activity at a low or routine level, 20 indicate interest at a medium level, and 10 indicate interest at a high level—supervisory or administrative. It is thus possible to score for levels of interest. Three other methods of scoring enable us to obtain three additional types of interest: (1) verbal, (2) manipulative, and (3) computational. The manual reports reliabilities of .82 to .93 for each field of interest. The norms are based on the records of 1,000 California children as well as of 954 male veterans.

The validation of the test is incomplete. Its occupational families correlate well with the Kuder divisions when they are really comparable. For example, personal-social correlates .60 with Kuder's social service; mechanical with Kuder's mechanical, .72; business with Kuder's clerical, .74; the sciences with Kuder's scientific, .80; and computational with Kuder's computational, .50.¹ The Lee-Thorpe inventory has little or no correlation with intelligence-test scores. The validation is incomplete because it has not been applied to persons engaged in a large variety of occupations. In short, the Lee-Thorpe inventory shows promise of being a very useful instrument for purposes of interviewing and with further study may develop into a very valuable interest inventory.

In considering which one of these four inventories to use, the matter of vocabulary load deserves some weight. One investigator made a study of the vocabulary load of seven inventories.² Our four were included in this seven. The results of the study are shown in the table at the top of page 435. It seems clear that from the standpoint of vocabulary load the inventories of Kuder and Lee-Thorpe are more suitable for the lower high school grades than those of Cleeton and Strong.

Table 16 contains a selected list of interest inventories. One of them such as Dunlap's Academic Preference Blank is suitable for younger children and is constructed for the purpose of discovering the interests

¹ These correlations are from Lindgren, Henry C., "A Study of Certain Aspects of the Lee-Thorpe Occupational Interest Inventory," *Journal of Educational Psychology* (1947) 38:353-362.

² Roeber, Edward C., "A Comparison of Seven Inventories with Respect to Word Usages," *Journal of Educational Research* (1948-49) 42:8-17.

| Inventory | Percentage of Different Words above the Level of Grade 9 |
|-----------------|---|
| Lee-Thorpe..... | 8.9-9.6 |
| Kuder..... | 10.6 |
| Cleeton..... | 14.9 |
| Strong..... | 16.1 |

of children in school subjects or areas of study. Garretson and Symonds Interest Questionnaire for High School Students helps us to distinguish between only three areas of interest: the academic, the technical, and the commercial. The others in the table have less value for our purposes.

DIRECT OBSERVATION OF THE INTERESTS OF CHILDREN AND STUDENTS

Direct observation is an ancillary and corroborative technique rather than a primary one. The motives of children, students, and adults are so complex that their actions are easily misinterpreted. And yet there are some possibilities here. The author¹ checked the records of children's interests obtained through a questionnaire by observing the children at their reading in public libraries. This was done (1) directly by observing the books which the children freely selected, and (2) indirectly by rating the blackness of the cards in the card catalogue as well as by recording the number of books worn out. In like manner records can be kept of the types of plays and games which individuals like at different seasons. Anecdotes also, if recorded at the time of the occurrence and accumulated from time to time, are valuable aids for discovering the range of children's interests.² For example, the author once observed a group of boys plan and construct a radio tower. They worked for the money to buy the materials and constructed the tower themselves. The record of such experiences forms a capital illustration of the anecdotal record.

INTERESTS THROUGH INFORMATION

A third procedure used in discovering interest is through measuring the information about a topic possessed by an individual and inferring from his information the amount of interest he has in it. On the one hand, you ask a subject to express his feeling toward an item in the terms of like, indifferent, dislike (L-I-D); on the other, you check his factual information. On the one hand, you ask him if he likes baseball; on the other, you question him to see if he knows what a "squeeze play" is, or a "fielder's choice." In the latter case you assume that if he knows most of the technical terms in baseball he would have a very great interest in it. Such an individual would have liked baseball, would have

¹ Jordan, A. M., *Children's Interests in Reading*. Chapel Hill: The University of North Carolina Press, 1926.

² Jarvie, L. L., and Mark Ellingson, *Handbook on the Anecdotal Behavior Journal*. Chicago: University of Chicago Press, 1940.

TABLE 16. INTEREST INVENTORIES

| Name | Grade | Types of scores | Reliability | Validity | Norms | Publisher | Time, minutes |
|--|----------------|---|--------------------------|--|--|---------------------------------------|---------------|
| Kuder Preference Record | 9-16 | | (See text) | | | Science Research Associates | 40-60 |
| Strong's Vocational Interest Blank | 10-16 | | .877 | (See text) | 39 occupations | Stanford University Press | 40 |
| Lee-Thorpe Occupational Interest Inventory | 9-11 and 12 up | | (See text) | | Norms for each type of score | California Test Bureau | 30-40 |
| Glaser-Maller Interest Values Inventory | 9-16 | Theoretic Aesthetic Social Economic | .91 .93 .92 .87 | Items kept which distinguished between the four types already known to be present in students | Norms for four types | Teachers College, Columbia University | 30 |
| Cleeton Vocational Interest Inventory | 9-16 | | (See text) | | Norms for the nine types | McKnight & McKnight | 45-55 |
| Brainard and Stewart Specific Interest Inventory | * | Subject's interest in a particular mode of expressing activity such as physical work, vocal expression, experimenting | .68 13-.94 | No objective data on reliability or validity of inventory. Question as to independence of separate dimensions: outdoor, scientific, experimentation, observation, and creative imagination | Norms for different modes of expressing activity | Psychological Corporation | 30-40 |

TABLE 16. INTEREST INVENTORIES (*Continued*)

| Name | Grade | Types of scores | Reliability | Validity | Norms | Publisher | Time, minutes |
|---|-------|--|---------------------|--|--|---------------------------------------|---------------|
| Garretson and Symonds Interest Questionnaire for High School Students | 8-10 | Academic Commercial Technical | .86 .925 .953 | Biserial correlation between selecting one curriculum rather than another. Predicted what curriculum a boy would choose | Norms for each of three types of curriculum | Teachers College, Columbia University | 30 |
| Dunlap Academic Preference Blank | 6-9 | Interest in eight areas of elementary school | .70 .83 | 90 words or phrases relate to special academic areas. Correlated with success in each area and with general intelligence | (1) Paragraph meaning, (2) word meaning, (3) history, (4) language usage, (5) geography, (6) literature, (7) arithmetic, (8) general achievement | World Book Company | 15 |

* Age: boys 10-16, men 16 and over, girls 10-16, women 16 and over.

read the rule book, would have had a pleasant glow when he discovered a new idea, and would have talked about it with others and as a consequence remembered it well. On the other hand, as a result of association with friends there might be an accumulation of facts on a topic in which a person has little interest.

The Information Test of Interests

In the construction of an information test which will reflect the interest of subjects, samples must be taken from the total number of experiences which an individual has had in that area. The number of items known will then indicate to some extent the amount of interest which the individual possesses in it. The most successful of these tests have been those dealing with mechanical and social interests. Such tests are illustrated by the Mechanical Interest Test which was the first part of the United States Army Mechanical Aptitude Test (1921). Other tests are (1) O'Rourke's Mechanical Aptitude Test, (2) Ream's Social Relation Test, (3) Burt's Agricultural Interest Test, (4) McHale's Vocational Interest Test for College Women, and (5) Toop's General Interest Test for Girls. The reliabilities of these tests have been found to be satisfactory. On the average, these coefficients have ranged from .89 and above. It is in the area of validity that they show their greatest weakness.

Validity

It is customary in establishing the validity of a test to compare the records of one test with records obtained from another test or from ratings of competent persons. These objective interest tests do correlate with each other. The coefficients range from .57 to .70, with a mean of .67.¹ This figure is not much below the intercorrelations of different intelligence tests. It is when these information tests are correlated with estimates of interests that their true validity is determined. In one case,² a direct comparison was made between estimated occupational interests and information scores with a resulting correlation of only .15. This is indeed a negligible relation. Experience with the Army Mechanical Interest Test indicated that probably this relation between measured

¹ Fryer, Douglas, *The Measurement of Interests*. New York: Henry Holt and Company, Inc., 1931. Chapter VIII of this book contains a competent treatment of the whole subject, much more extended than that in the present volume. In Fryer's text have been gathered the correlations on reliability and validity of these objective interest tests.

² McHale, Kathryn, "An Information Test of Interests," *Psychological Clinic* (1930) 19:53-58.

and estimated interests was a little higher than .15, perhaps .23 or .24. This relation is a low one at best and indicates that these two attempts to get at interests were emphasizing different aspects of the problem.

As in the case of subjective interests the question arises as to the relation between the scores on tests of objective interests and on measures of achievement. The average coefficient of correlation between objective interest scores and scores on such measuring instruments as Stenquist Assembly Test or Stenquist Picture Tests, which are measures of mechanical ability, was above .40. It is clear that interest and success while being correlated are sufficiently unlike to demand different types of measuring instruments.

Intelligence tests and objective interest tests vary in the size of the coefficients of correlation. The cause of this variation lies in the type of interest being measured. If the interests are social, such as are measured in Ream's Social Relations Test, then the correlation is marked. If, on the other hand, the interest measure is an indicator of mechanical interests, the coefficient is much lower. In the former case the coefficients range around .60; in the latter, around .40. Tests of general information have, since the advent of Army Alpha, stood up well as measures of intelligence. Their correlations with other indicators of intelligence have been as high as most other forms used in the measurement of intelligence. It is thus indicated that a score on an objective test of interest is measuring intelligence in part. This is an anticipated finding, since the acquisition of information has long been recognized as the result of the joint influence of interest and intelligence.

From these considerations it seems evident that the scores on a test of objective interest are composed of much more than the mere interest that an individual has in that area. Intelligence and past experience, whether of interest or not, are additional factors. At any rate the scores are not clear-cut measures of interest. Because of this ambiguity in the meaning of the scores, objective tests of interests have never gained the popularity that subjective measures have. If some technique could be found which would separate the interest factor from the others these objective tests of interest would forge rapidly to the front. Objective tests of interest have tended to be merged with tests of aptitude and achievement, leaving the interest factors to be measured by the subjective tests.

USES OF INTEREST INVENTORIES

Scores from a well-administered and unprejudiced taking of interest inventories are of use in four areas: (1) they help an individual assess his own interests, (2) they are useful to the counselor and the pupil in educational guidance, (3) they help the student in his choice of an

occupation, and (4) they aid the teacher in motivating and expanding the work of the classroom.

In the first place, the study and discovery of a student's interests are valuable in his personal, educational, and occupational or vocational development. The taking and study of such an interest inventory as has been described in this chapter establishes in the individual the habit of studying his own personality traits objectively. He finds, for example, that his real interests are different from those advised by his parents or wished for by his teacher. It brings him face to face with his own choice of occupation. This in itself is a sobering thought. If the teacher or counselor goes over with him his strong and weak interests he may free himself of his inhibitions and talk out his most intimate interests. This may lead to a discussion of the aptitude requirements of various occupations and of whether or not he possesses them. Thus the student is helped to think about the direction of his life and to appraise his own traits carefully and objectively. It may help the undecided to decide upon an occupation which, while it might not be permanent, would give some direction to a student's growth.

In the second place, once the direction of a student's life is pretty well agreed upon there follows a discussion of the subjects of instruction which give material aid in the fulfillment of his desire. In what school subjects is he now interested, what significant ones has he not taken, what should be his curriculum in the light of these interests? It is thus possible to guide such a student into subjects which furnish him most interest. However, the claims of interest should not be too weighty since interest is not closely related with either aptitude or achievement. But other things being equal, the student should be encouraged to take those courses in which he has a genuine interest.

In the third place, the purpose of tests in guidance is to help the student find an area of occupations which he might successfully enter. The type of occupation selected in this manner might affect his determination to go on to college or to take further training. If he made up his own mind in favor of an occupation based on his own interests he might be more willing to work harder in preparation for it. Moreover, he would subsequently find in that occupation workers who had the same interest with himself and would add to his sum total of happiness. As Cleeton says in his manual¹ that this procedure would help a student get into an occupation where he would have "fewest personal handicaps and the greatest personal satisfaction."

In the fourth place, many uses can be made of children's interests within the classroom. Interests in radio programs, reading, moving

¹ Cleeton, Glen U., *Manual of Directions for Cleeton Vocational Interest Inventory*, p. 8. Bloomington, Ill.: McKnight and McKnight, 1943.

pictures, or events of the day can be used along with inventoried interests to motivate children's learning. Projects based on such interests and involving activities growing out of them may develop meanings and expand horizons which otherwise might have remained little understood and narrow. To a child interested in adventure such books as *Treasure Island* or *Call of the Wild* are a godsend. It is well for teachers to know the interests of their students, for in answering their questions and directing their activities a type of education may be developed which will continue long after the course is completed.

RELATION OF INVENTORIED INTERESTS TO OTHER TRAITS

The amount of relationship of interests to (1) measures of achievement, (2) measures of general intelligence, and (3) measures of special aptitudes is of considerable importance.

The coefficient of correlation computed between school marks and interest or between achievement tests and interest is not high. In general this relationship is represented by a coefficient of correlation between .00 and .40. Garretson and Symonds¹ report no resemblance between commercial interests and commercial grades ($r = .00$), but a slightly higher coefficient ($r = .29$) between technical interests and grades in technical subjects. Correlations between the Kuder Preference Record and school achievement as measured by standard tests have been reported somewhat higher. One study² showed that the coefficients between interest in science and general science achievement was .42 for women and .32 for men, while the corresponding coefficients between interest in literature and achievement in literature was .33 for women and .40 for men. Most of the coefficients computed from Strong's Vocational Interest Blank and achievement in school have been done at the college level and, in general, are slightly lower than those here reported.

The relation between measured intelligence and inventoried interests resembles rather closely that between interest and achievement. In one study (Kornhauser, 1929)³ the reported correlation between the Kornhauser General Interest Inventory and intelligence was .29. When

¹ Symonds, P. M., and O. K. Garretson, *Interest Questionnaire for High School Students*. New York: Bureau of Publications, Teachers College, Columbia University, 1930.

² Triggs, F. O., "A Study of the Relation of Kuder Preference Record Scores to Various Other Measures," *Educational and Psychological Measurement* (1943) 3:341-354.

³ Kornhauser, A. W., "Results from a Quantitative Questionnaire on Likes and Dislikes with a Group of College Freshmen," *Journal of Applied Psychology* (1929) 11:85-94.

Primary Mental Abilities¹ were correlated with the different interest scores (the Kuder Preference Record) the correlations when 512 university freshmen were used as subjects were low with but one exception. Computational interest had a present but low correlation (.39) with number ability.²

Special aptitudes and scores from interest inventories are inclined to be only loosely related. In one extensive study,³ which included subjects from grade 7 to freshmen in college, scores from an interest inventory (Interest Analysis Blank for Boys) correlated from .00 to .35 with measures of mechanical abilities. With the Minnesota Spatial Relations Test the interest scores correlated from .09 to .30; and with the Minnesota Assembly Test and the Minnesota Paper Form Board the coefficients were no different. Finally, when the Mechanical Abilities Battery was used the coefficients with the Interest Analysis Blank for Boys varied from .00 to .35. It is clear that one cannot depend on mechanical interests to predict tested mechanical abilities.

From the evidence presented here, but much more from the total available evidence, it may be clearly inferred that interests are separate abilities. One can predict from interest scores neither achievement nor intelligence nor special aptitude. Measures of these last must be gathered from separate tests. Since these great areas of interests, achievement, intelligence, and special aptitudes are separate, some provision must be made to bring them all together so that one can consider more effectively the whole child. Such an attempt will now be described.

In the city of Philadelphia an experiment in guidance was undertaken with several classes. In this undertaking, tests of intelligence (the Chicago tests), tests of school achievement, the Minnesota Paper Form Board, and the Kuder Preference Record were used. The pupils, furnished with graph paper and scores from the various tests, were taught to enter them correctly on the graph. Each child then considered his own abilities as assessed by the measuring instruments and appraised them in the light of his vocational choices. Sometimes the parents, the child, and the counselor considered them together. Figure 35 along with its description shows concretely how these data are used.

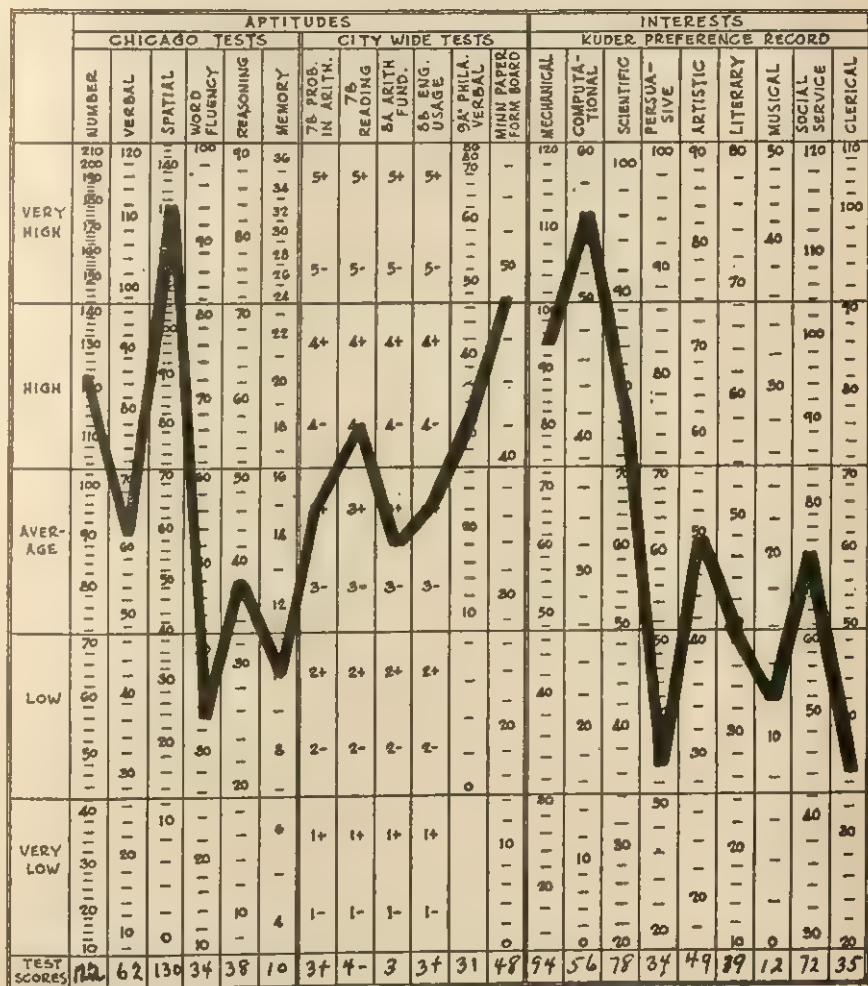
In this last illustration is indicated the very best use to which tests can be put. In no case should the individual be lost in the testing

¹ Adkins, D. C., and G. F. Kuder, "The Relation between Primary Mental Abilities and Activity Preferences," *Psychometrika* (1940) 5:251-262.

² See Super, Donald E., *Appraising Vocational Fitness*, Chaps. XVII, XVIII. New York. Harper & Brothers. These chapters give a much more complete treatment of these matters.

³ Hubbard, R. M., "A Measurement of Mechanical Interests," *Journal of Genetic Psychology* (1928) 35:229-252.

PROFILE CHART SELF-APPRAISAL PROGRAM OF GUIDANCE IN THE NORTH JUNIOR HIGH SCHOOL
 PUPIL'S NAME BEN WILLARD DATE OF FIRST ENTRY SEPT 1945
 RESIDENCE 1846 BROAD ST. ADVISER MR. STITH
 CAREER PLANS: 1- DRAFTSMAN 2- ENGINEER
 TENTH GRADE SELECTIONS: SCHOOL PARK HIGH CURRICULUM MECH. ARTS



FORM H 149-PROFILE, SELF-APPRAISAL, JUNIOR HIGH-SCHOOL, DISTRICT OF PHILADELPHIA

FIG. 35. From Self-appraisal Program of Guidance in the Junior High School. (By permission of Louis P. Hoyer, superintendent, School District of Philadelphia, 1947.)

shuffle. It is he on whom the purer light of test results is focused. If he is not profited the whole testing process is nothing but a tinkling cymbal.

Figure 35 may be interpreted as follows:

The first chart describes a boy with highest aptitude scores in number and in spatial thinking. His chief interests are in mechanical, computational, and scientific fields. He seems to have little aptitude in word fluency and, likewise, little interest in persuasive, musical, and clerical activities. High aptitudes in number and spatial thinking forecast probable success in mechanical fields. Strong interests in mechanical and scientific areas on this chart would seem to indicate that this boy would be happy as well as successful in work where speaking and writing are not important.

This profile is helpful to the boy's parents. They can see clearly that he could become a skilled mechanic, a technician in industry, or, with college training, he might become a mechanical engineer.

Several opportunities are open to him at present. One is a curriculum in his neighborhood senior high school that will include mechanical construction. Another is the machine construction, drafting, or other programs offered by the vocational-technical schools.

Future opportunities for additional training should be considered. After graduation from a regular three-year vocational-technical school program, one or two years of advanced work would enable him to qualify for the vocational-technical diploma and obtain employment as a technician in industry. Or, while employed at his trade, he could attend the Standard Evening High School if any additional units were needed for entrance into the engineering college of his choice.¹

SUMMARY

Three techniques have been tried out to discover interests: (1) direct questioning, (2) observation, and (3) objective tests of information. Of these, the direct questioning of the subjects has been the most successful. The weakness of direct questioning has been realized: (1) actual lying about the items to make an impression, (2) failure to generalize correctly from experienced events, and (3) lack of information about the item in question. In spite of such shortcomings, questionnaires have proved to be both reliable and valid. Their scores vary in specificity from an interest score in a well-known occupation such as a certified public accountant to the designation of certain areas of interest such as artistic or mechanical. Observation of the subjects' activities have

¹This description of Fig. 35 appears in *Self-appraisal Program of Guidance in the Junior High School*. School District of Philadelphia, 1947.

proved valuable only as evidence supplementary and confirmatory to the other techniques. The objective tests of interests at first glance would appear to be the most promising of all the techniques thus far described. However, because information in any form is closely correlated with both intelligence and achievement, difficulties have arisen which are thus far insurmountable. Promising beginnings in this area of objective testing of interests have been never quite fulfilled.

The uses of these measures of interest have been widespread. They have been found useful in aiding the classroom teacher to direct the interests already present as well as in assisting the counselor to help the student select a program of studies. In the area of vocational counseling these interest scores have proved useful in getting a willing subject to view objectively the types of interests which he actually possesses. Taking and scoring such an inventory encourages a subject to assume an objective attitude toward his own interests. Finally, the taking of such inventories aids in narrowing the field of occupations which the student might enter. He finds, for example, that his interests are clearly mechanical, a fact which definitely limits the vocations to be considered.

QUESTIONS AND EXERCISES

1. Why is it said that the main purpose in the measurement of interests is for guidance?
2. How are motives and interests related? How different?
3. Describe and evaluate three principal methods used in the discovery of interests.
4. Why has not the amount of information in any area reflected the amount of interest present?
5. What are three principal sources of errors to be considered in using the interest questionnaire?
6. Describe the process used in validating interest questionnaires.
7. What are the leading features of (a) Strong's Vocational Interest Blank, (b) Cleeton's Vocational Interest Inventory, (c) the Kuder Preference Record, and (d) the Lee-Thorpe Occupational Interest Inventory.
8. Explain precisely how these inventories can be used by the teacher and by the counselor.
9. a. Make a table which includes the divisions of interest obtained from scoring (1) Cleeton's Vocational Interest Inventory, (2) Kuder Preference Record, and (3) Lee-Thorpe Occupational Interest Inventory.
b. Which seems to you the most useful arrangement? Why?
10. What are the conclusions concerning the permanence of interest in (a) the elementary school, (b) the high school, and (c) the college?
11. Discuss the uses to which interest inventories may be put.

BIBLIOGRAPHY

Books

- FRYER, DOUGLAS: *The Measurement of Interests*. New York: Henry Holt and Company, Inc., 1931.
- GREENE, EDWARD B.: *Measurements*

of Human Behavior, Chap. XV. New York: The Odyssey Press, Inc., 1941.

JORDAN, A. M.: *Children's Interests in Reading*. Chapel Hill: The University of North Carolina Press, 1926.

REMMERS, H. H., and N. L. GAGE: *Educational Measurement and Evaluation*, pp. 407-425. New York: Harper & Brothers, 1943.

SMITH, EUGENE R., RALPH W. TYLER, et al.: *Appraising and Recording Student Progress*, pp. 358-402. New York: Harper & Brothers, 1942.

STRONG, E. K., JR.: *Vocational Interests of Men and Women*. Stanford University, Calif.: Stanford University Press, 1943.

SUPER, DONALD E.: *Appraising Vocational Fitness*, Chaps. XVI, XVII, XVIII. New York: Harper & Brothers, 1949.

Articles in Journals, Manuals

ADKINS, D. C., and G. F. KUDER: "The Relation between Primary Mental Abilities and Activity Preferences," *Psychometrika* (1940) 5:251-262.

CANNING, L. B., KATHERINE VAN F. TAYLOR, and H. D. CARTER: "Permanence of Vocational Interests of High School Boys," *Journal of Educational Psychology* (1941) 32:487-493.

CARTER, H. D., K. V. F. TAYLOR, and L. B. CANNING: "Vocational Choices and Interest Test Scores of High School Students," *Journal of Psychology* (1941) 11:297-306.

CLEETON, GLEN U.: *Manual of Directions for Cleeton Vocational Interest Inventory*. Bloomington, Ill.: McKnight and McKnight, 1943.

FRANSDEN, ARDEN: "Appraisal of Interest in Guidance," *Journal of Educational Research* (1945-1946) 39:1-12.

HUBBARD, R. M.: "Measurement of Mechanical Interests," *Journal of Genetic Psychology* (1928) 35:229-252.

JARVIE, L. L., and MARK ELLINGSON: *Handbook on the Anecdotal Behavior Journal*. Chicago: University of Chicago Press, 1940.

KORNEHAUSER, A. W.: "Results from a Quantitative Questionnaire on Likes and

Dislikes with a Group of College Freshmen," *Journal of Applied Psychology* (1929) 11:85-94.

KUDER, G. F.: *Manual to the Kuder Preference Record*. Chicago: Science Research Associates, 1939, 1946.

LINDGREN, HENRY C.: "A Study of Certain Aspects of the Lee-Thorpe Occupational Interest Inventory," *Journal of Educational Psychology* (1947) 38:353-362.

MCMALE, KATHRYN: "An Information Test of Interests," *Psychological Clinic* (1930) 19:53-58.

ROEBER, EDWARD C.: "A Comparison of Seven Interest Inventories with Respect to Word Usage," *Journal of Educational Research* (1948-1949) 42:8-17.

Self-appraisal Program of Guidance in the Junior High School. District of Philadelphia, 1947.

STRONG, EDWARD K., JR.: "Weighted vs. Unit Scores," *Journal of Educational Psychology* (1945) 36:193-216.

TRAXLER, A. E., and WILLIAM C. MCCALL: "Some Data on the Kuder Preference Record," *Educational and Psychological Measurement* (1941) 1:253-268.

TRIGGS, FRANCES ORALIND: "A Study of the Relation of Kuder Preference Record Scores to Various Other Measures," *Educational and Psychological Measurement* (1943) 3:341-354.

—: "A Further Comparison of Interest Measurement by the Kuder Preference Record and the Strong Vocational Interest Blank for Men," *Journal of Educational Research* (1943-1944) 37:538, 544; also (1944-1945) 38:193-200.

WITTENBORN, J. R., FRANCES ORALIND TRIGGS, and DANIEL D. FEDER: "A Comparison of Interest Measurement by the Kuder Preference Record and the Strong Vocational Interest Blanks for Men and Women," *Educational and Psychological Measurement* (1943) 3:239-257.

CHAPTER 17

Measurement of Attitudes

Attitudes and interests determine pretty largely the direction of behavior. Even more than knowledge, attitudes affect action. In the realm of alcoholic consumption, persons learn all about the evil effects of alcohol and then drink large quantities of it. If, however, an emotionally toned attitude is built up against it or in favor of it, action follows much more certainly. In a great many areas of life is this true. In the fields of government, economics, labor relations, taxation for schools, militarism, internationalism, race relations, social relations, and in many other relations, attitudes play a dominant part in determining action. If, then, attitudes of adults are so important, why should they not be of the greatest importance in the schools? The answer is, of course, that they are and that definite evidence of their development should be made available.

Measurement, if well developed, could help in providing attested evidence of the presence of desired amounts of an attitude if the attitude had already been carefully described as one of the outcomes of instruction. Unfortunately agreed-upon lists of attitudes desirable for attainment in school have not been made, and as a result, development of measuring scales and instruments directly useful in the school situation has been delayed. Another cause for the confusion in this area has been the variety of definitions of attitudes developed by competent psychologists. In one case psychologists define an attitude in rather general terms as "a more or less emotionalized tendency organized through experience, to react positively or negatively toward (for or against) a psychological object" (Remmers and Gage). Here all attitudes would involve some feeling *for* or *against* a psychological object. A psychological object would be one which aroused reactions in individuals. One can readily see that this might be a latent tendency such as the one to be kind to dumb animals or to aid those in distress, but it also might mean a belief in some movement—for example, that for government housing—or a position taken in regard to the democratic way of life. One other feature of this definition must not be forgotten; it must be organized through experience. Attitudes as we generally study their acquisition are certainly learned and organized through experience.

Little white children in the South are not born with attitudes toward the Negro but gather them from their personal experience.

Let us look at one or two other definitions of attitudes. An attitude is a "set or disposition to act toward an object according to its characteristics as far as we are acquainted with them" (Woodworth). In this definition "set" or "disposition" substitutes for "emotionalized tendency." It too, emphasizes environment in the phrase "as far as we are acquainted with them." "Object" would also be a psychological object. A second definition also commands our attention: "an enduring acquired predisposition to react in a characteristic way, usually favorably, or unfavorably, toward a given type of person, object, situation, or ideal" (Dashiell). Notice the emphasis here on the word "enduring." Unless the experience were enduring it would hardly be called an attitude. Otherwise we would think of it as a mood or a temporary emotion. "Predisposition" here corresponds to "tendency" in Remmers and Gage definition and to "set" or "disposition" in Woodworth's. Dashiell's "favorably or unfavorably" corresponds to "positively or negatively" of Remmers and Gage.

Out of these definitions and their discussion come some of the leading characteristics of an attitude:

1. An attitude is essentially a *set* or *disposition* which is also described as a predisposition or tendency.
2. There is almost always a *feeling tone* to act favorably or unfavorably, positively or negatively toward an object.
3. The attitude is a *result of experience*.
4. The set or disposition is directed *toward some psychological object* such as a person, a situation, an institution, a race, or an ideal.
5. It is *enduring*.

It is thus seen that attitude is a broad term and that it would be quite impossible to obtain scales and measures for all attitudes even if it were desired. The problem for the school is to define a number of the most desirable attitudes so specifically and clearly that measurement will be possible.

THE LEARNING OF ATTITUDES

Attitudes are learned much as are other experiences. Sometimes an attitude is acquired through one dramatic experience. A young boy wishing to be manly takes a chew of tobacco. He is made deathly sick and as a result forms an attitude toward tobacco which may last for years. In the second place, attitudes may be acquired through several repetitions of a similar experience; such is the case with our attitude toward Russia which now seems to be well formed. Now and then an experience is simply absorbed from the environment in such subtle

ways that it is difficult to describe. Note the attitude of a child toward labor unions if he has been reared in a home of a manager of a large business. Again, the attitude formed may simply be produced by a process of integration, as when a student who has failed one foreign language dislikes all of them. The case of a boy comes to mind who learned to dislike teachers in general because his music teacher lost her temper to such an extent that she slapped him in the face. This last example had rather ludicrous repercussions because the lad in question organized his comrades to sing loudly and lustily offkey whenever his teacher wanted them to sing especially well. An 8-year-old white boy living on a farm admired greatly a Negro carpenter who used to come over to the home place to build a shed, repair a roof, or mend whatever was broken. The boy used to assist the carpenter and enjoyed thoroughly the days when the carpenter came. He even called the man "Mr. Savage." His elders, on hearing the boy say "Mr. Savage," said, "You mustn't call him 'Mister.' He is a nigger." This was said with such emphasis that the lad knew his mistake must not be repeated. These learning processes are worthy of consideration because they apply both when attitudes are to be learned and when they are to be changed.

ATTITUDES WITH WHICH MEASUREMENT IS CONCERNED

Since attitudes may be developed toward almost any object, individual, institution, or race it would be manifestly impossible to develop measuring scales for all of them. The reasonable procedure seems, then, to select some of the most far-reaching ones for both instruction and measurement. There is some danger here of making the categories so broad that their specific applications are blurred. One investigator¹ attempted to discover what objects an unselected population regarded as socially significant. He found 238 objects which two psychologists classified into eight categories:

1. Personality
2. Education
3. Economic activities
4. Family
5. Government
6. Social problems
7. Recreation and exercise
8. Religion

It is clear that such broad areas must be broken down into smaller more clearly defined units before they could possibly be of much value for the educative process.

¹ Horne, E. Porter, "Socially Significant Attitude Objects," *Studies in Higher Education*, XXXI, *Bulletin of Purdue University* (1936) 37:117-126.

A second attempt to list some of the educable attitudes holds out more promise of success.

In attempting to discover areas of social belief which were of primary concern to the school one set of investigators (Smith and Tyler, 1942) asked students, principals, teachers, and parents to suggest the areas of social beliefs in which they were interested. The following areas of social issues seemed most important: "democracy political and economic, the role of the machine and invention in contemporary civilization, consumer problems, use of natural resources, labor, unemployment, housing, nationalism and internationalism, war and peace, school life, religion, and family." The authors chose the areas of social issues for developing their instruments for measuring attitudes.

One illustration from this study will give concreteness to the discussion. The area in question was concerned with beliefs about school life. In analyzing this area of social beliefs students were asked to write essays on "democracy in my school." The lists of areas collected from this procedure were added to those of teachers and staff and sent around with illustrative statements of issues in each area to teachers in several schools. These teachers were asked to evaluate the suggested areas of social attitudes and to add others. This procedure resulted in six major areas "school government, curriculum, grades and awards, school spirit, pupil-teacher relations, and group life." When these were analyzed into subareas, many concrete problems were raised such as (1) whether the opinions of students should be solicited and heeded in buying new books, (2) whether students from wealthier families should be put in the same home rooms with those from poorer families, and (3) whether it is better for the teacher to decide what is to be studied in class or for the students to plan their work themselves.

If we compare the two lists of attitude objects on page 449 and on the present page, it is clear that these two sets of investigators had different things in mind because there are only a few items that occur in both lists. After we pass "home and family" and "education" the rest of the items differ in name. This fact of disagreement illustrates the difficulty in developing good attitude scales because all good measurement depends upon a clear-cut definition of the goal or objective sought for. Measurement, then, consists of stating how far along the road to the objective an individual has gone. Until the attitude objects are clearly defined the measurement of attitudes must remain in the experimental stage.

THE MEASUREMENT OF ATTITUDES

Since it is impossible to secure an agreed-upon list of attitude objects, the only course left to the student of education is to select from the

extant scales those scales which may be of value to the work of the school. Types of scales, inventories, and other techniques will now be presented and evaluated.

In measuring an attitude, the ideal situation would be to have a series of unambiguous statements placed at equal intervals on a scale ranging from absolute approval to absolute disapproval. Each statement would be chosen because it expressed clearly and certainly a defined position on the scale. A person wishing to discover his own attitude could then check the items or statements with which he agreed, add up the positional points and divide this sum by their number, thus obtaining his position on the scale.

This ideal of equal units has not been attained. The nearest approach to it are the Thurstone scales constructed upon the *principle of equal-appearing units*. The units are equal because they appear equal to the competent persons who sorted the statements into defined piles. In constructing the scale on the attitude toward the church (Thurstone and Chave, 1929) 130 statements were collected which reflected varying degrees of friendliness or unfriendliness toward the church. These statements were sorted into 11 piles by 300 sorters. Eleven master slips, designated A to K, were placed upon a table upon which the statements were to be placed by the sorters. The positions of three of the master letters were defined. In Pile A were placed those statements which expressed highest appreciation of the church; in Pile K, those statements which expressed the strongest depreciation of the church; and in Pile F, only neutral expressions. The remaining letters in the series were left undefined. Certain criteria were used to prevent statements from being ambiguous. Let us take one statement about the church: "I have seen no value in the church." If this statement had been placed at F, G, H, and K by a substantial number of sorters with the largest number at H, let us say, it would not have been accepted as an item. It would have been ambiguous. Thurstone checked this matter of ambiguity by subtracting the 25th percentile from the 75th percentile. A much better situation would arise did the great majority of sorters place the item at H with a very small number placing it at G or K. As a matter of fact this particular item was not at all ambiguous and was placed at 9.9, just about at H. In this manner a series of statements was drawn up ranging from A to K or from 1 to 11 and expressing different degrees of belief in the church from extreme belief to extreme disbelief, the statements being placed at equal-appearing intervals.

In like manner was constructed the scale of Communism. This scale is comparatively easy to use. One simply checks the items in which he believes. These items, although arranged irregularly on this page, may be thought of as arranged in a series according to their scale values. The

ATTITUDE TOWARD COMMUNISM, SCALE NO. 6, FORM A,¹

(Prepared by L. L. Thurstone)

Put a check mark (✓) if you agree with the statement

Put a cross (X) if you disagree with the statement

1. Both the evils and the benefits of communism are greatly exaggerated.
3. The whole world must be converted to communism.
5. Communism is a much more radical change than we should undertake
7. Give Russia another twenty years or so and you'll see that communism can be made to work.
9. Communism should be established by force if necessary.
11. I am not worrying, for I don't think there's the slightest chance that communism will be adopted here.
13. Communism is the solution to our present economic problems.
15. The ideals of communism are worth working for.
17. The whole communistic scheme is unsound.
19. We should not reject communism until it has been given a longer trial.

median or *average score* of the items checked is then used as the best representative of the subject's position. One may also use his *range of scores* as another measure and the *one statement which most nearly represents his position* as the third measure.

There are many more scales constructed by Thurstone himself and under his leadership which are of great interest to school people: Here are 17 scales of interest to high school teachers:

1. War (D. D. Droba)
2. The Negro (E. D. Hinckley)
3. The law (D. Katz)
4. The Germans (R. C. Peterson)
5. The Constitution of the United States (A. C. Rosander and L. L. Thurstone)
6. Prohibition (H. H. Smith and L. L. Thurstone)
7. Communism (L. L. Thurstone)
8. Monroe Doctrine (L. L. Thurstone)
9. Freedom of speech (L. L. Thurstone)
10. Honesty in public office (L. L. Thurstone)
11. Public ownership (L. L. Thurstone)
12. Unions (L. L. Thurstone)
13. The treatment of criminals (C. K. A. Wang and L. L. Thurstone)
14. The movies (L. L. Thurstone)
15. German war guilt (L. L. Thurstone)
16. Divorce (L. L. Thurstone)
17. The Chinese (R. C. Peterson)

No one doubts that the attitude scales are good or that they are as rigorously constructed as any known at the present time. One weakness

¹ By permission of University of Chicago Press.

appears. The registering of attitudes is desired toward hundreds of psychological objects. To construct single scales for each such object would require more work than can be afforded. Is it not possible to develop a sort of general scale which could be used toward several objects under certain conditions?

Let us first consider the Bogardus Scale of Social Distance, which is in the form of a rating scale. It indicates the degree of closeness to which an individual is willing to admit members of another race. The scale is as follows: (1) to close kinship by marriage, (2) to my club as personal chums, (3) to my street as neighbors, (4) to employment in my occupation in my country, (5) to citizenship in my country, (6) as visitors only in my country, (7) would exclude from my country. These headings may appear at the top of a page and under each heading there may be written the appropriate number, such as:

| | | | | | | | |
|-----------|---|---|---|---|---|---|---|
| Canadians | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Germans | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

The attitude toward each nationality may be expressed by simply encircling one number. While this is only a rough rating scale, it does show the general attitude toward a race with some degree of consistency. The scale may also be used to express attitudes toward various religious denominations, liberals, agnostics, Socialists, or Communists. It has the advantage of ease of administration and of showing the general attitude of the rater. It lacks the precision of the Thurstone techniques just now described or even of the master scales whose consideration is now entered upon.

A second attempt to provide a more general scale appears in the Remmers Master Attitude Scales¹ which were developed according to the equal-appearing units of Thurstone. There is no difference in the principle of construction. The difference lies pretty largely in the generality of the statements. To be satisfactory a scale for *any* nationality would have to be stated very broadly so as to include Russian and German, French and Jugoslavs, etc.

Suppose we consider two of these scales which might be used for the same purpose. Let the problem be the expression of attitude toward the Negro race. We might use Hinckley's Attitude toward the Negro, 'Form A,' constructed after the principles of Thurstone, or we might use Grice's Generalized Scale Designed to Measure Attitudes toward Defined Groups, constructed according to the principles of Remmers.

¹ Remmers, H. H., and N. L. Gage, *Educational Measurement and Evaluation* New York: Harper & Brothers, 1943.

The following are 7 of the 16 statements from Hinckley's: Attitude toward the Negro.¹

- () 1. The difference between the black and white races is not one of mere degree, but of kind.
- () 3. No Negro should hold an office of trust, honor or profit.
- () 8. Inherently, the Negro and the white man are equal.
- () 9. The inability of the Negroes to develop outstanding leaders dooms them to a low place in society.
- () 11. After you have educated the Negro to the level of the white man, there will still be an impassable gulf between them.
- () 13. The Negro is by no means fit for social equality with the commonest white people.
- () 15. It is possible for the white and Negro races to be brothers in Christ without becoming brothers-in-law.

The following are the odd-numbered items from Grice's Generalized Scale Designed to Measure Attitudes toward Defined Groups, Form A (scale value in parenthesis):

- 1. Show a high rate of efficiency in anything they attempt (10.9)
- 3. Are mentally strong (10.0)
- 5. Are very patriotic (9.8)
- 7. Are noted for their industry (9.3)
- 9. Are a tactful group of people (9.0)
- 11. I would be willing to trust these people (8.8)
- 13. Command the respect of any group (8.5)
- 15. Are of a self-sacrificing nature (8.2)
- 17. Should be permitted to enter the country as immigrants (8.0)
- 19. Are a God-fearing group (7.7)
- 21. Are highly emotional (6.0)
- 23. Are superstitious (4.6)
- 25. Are unimaginative (4.2)
- 27. So far as I am concerned this group can stay in their native country (3.2)
- 29. Are frivolous (3.0)
- 31. I am not in sympathy with these people (2.5)
- 33. Are tactless (2.3)
- 35. Are despised by the better groups (1.9)
- 37. Belong to a low social level (1.6)
- 39. Should not be permitted to associate with other groups (1.4)
- 41. All members of this group should be deported from this country (1.2)
- 43. Respect only brute force (.9)
- 45. Are our worst citizens (.7)

By observing the Hinckley scale one is immediately aware that the statements are not arranged either in an increasingly favorable or

¹ Permission for using parts of Hinckley's and Grice's scales from H. H. Remmers, Purdue University, Lafayette, Ind.

increasingly unfavorable manner. One simply checks the statements believed in, obtains their scale value from another sheet, and averages the scale values. In the Grice scale the statements are arranged from extremely favorable to extremely unfavorable. The scale values are immediately before the user. The median score of the items checked is used to make the scoring very simple and rapid. The general scale does lose something of the concreteness of the particular scale. The general scale, moreover, has items which simply are not applicable in some cases such as No. 17 which refers to immigration, which is not a problem in the case of Negroes.

But after all the proof of the pudding is in the eating. If these scales furnish instruments which register faithfully a subject's attitude toward a race they are valuable whatever deadwood they contain. Grice applied the general scale and two of Thurstone's particular scales to the problem of obtaining attitude toward the Negroes and the Chinese. Thurstone's scale reliabilities varied around .87; that of the general scale was .84. When the records from the general scale were correlated with the records from the particular scale the coefficients ran from .58 to .75. Thus the evidence is in favor of using either test to register an attitude. If this is true in general, there is a great gain in economy in using the general test because it can be used in a great many situations. Consider Miller's scale on which one could express on one scale his attitude toward any of the 25,000 vocations listed in the United States. Remmers and his students have constructed 11 scales, known as Master Attitude Scales, by means of which a subject may measure his attitude toward the following:

1. Any disciplinary procedure (V. R. Clause)
2. Any elementary teacher (M. Amalara)
3. Any home making activity (B. K. Vogel)
4. Any play (M. Dimmit)
5. Any practice (H. W. Bues)
6. Any proposed social action (D. M. Thomas)
7. Any racial or national group (H. H. Grice)
8. Any school subject (E. B. Silance)
9. Any social institution (I. B. Kelly)
10. Any teacher (L. B. Hoshaw)
11. Any vocation (H. E. Miller)

Another attempt at measuring attitudes or beliefs is the Test of Beliefs on Social Issues.¹ These are most interesting because they grew up out of the school experiences of high school students and represent

¹ Procedure described in Smith, Eugene R., Ralph W. Tyler, *et al.*, *Appraising and Recording Student Progress*, pp. 209-229. New York: Harper & Brothers, 1942. Items by permission of Harper & Brothers.

to an extent their own experiences. It is reported: "In several cases both students and parents as well as teachers participated in this exploration—samples of student writing were analyzed, as were their choices of 'research' topics and free reading." Forms were developed both for the senior high school and for the junior high school level. The construction of this test differed from either the Thurstone or Remmers methods. The instrument consists of 200 statements "classified under the following areas of issues: democracy, economic relations, labor and unemployment, race, nationalism, and militarism."

Students respond to each issue by agreement, disagreement, or uncertainty. "The statements are arranged in random order and are presented to the students in two sections given at different times. For each statement in the first section there is a statement in the second section representing the opposite point of view." Two illustrations of the way the opposite items are phrased and of how they appear in different forms will be presented. The first pair deals with labor and unemployment:

- 4.21 14. Most workers who are unable to provide for themselves during a period of unemployment have been too shiftless to save. Agree, Disagree, Uncertain.
- 4.31 104. The wages of most workers are so low that it is impossible for them to save enough money to support themselves during periods of unemployment. Agree, Disagree, Uncertain.

The second pair deals with nationalism:

- 4.21 79. Our government ought to protect American business interests in foreign countries even if it involves using our army and navy. Agree, Disagree, Uncertain.
- 4.31 189. Our government should not risk a war to protect American business interests in foreign countries. Agree, Disagree, Uncertain.

The validity of these tests was checked by measuring the attitudes stated on the test against the opinions of the teachers regarding the students' attitudes. Furthermore, 30 students were interviewed and responded to oral questions similar to the ones on the written test. There was a fair consistency between the oral and written expressions of attitudes. The reliability estimated by the Kuder-Richardson formula from a population of 600 students, attending 14 schools and extending over a range from grade 9 to grade 12, ranged from .79 to .96. Reliability was also computed for liberalism, conservatism, and uncertainty. For total score on liberalism the coefficient of reliability was .95; uncertainty, .96; and on conservatism, .93. Unfortunately these scales have not been standardized.

An interesting attempt to discover the rise and growth of attitudes toward the Negro was made through the use of pictures.¹ Altogether

¹Hartley, Eugene L., "The Development of Attitude toward the Negro," *Archives of Psychology* (1936) No. 194, p. 47. Items by permission.

three tests were constructed. The first two tests might be answered directly from the pictures in Fig. 36. These pictures were judged by competent persons who had had much experience with both Negroes and whites to be both pleasing and typical of the races studied and

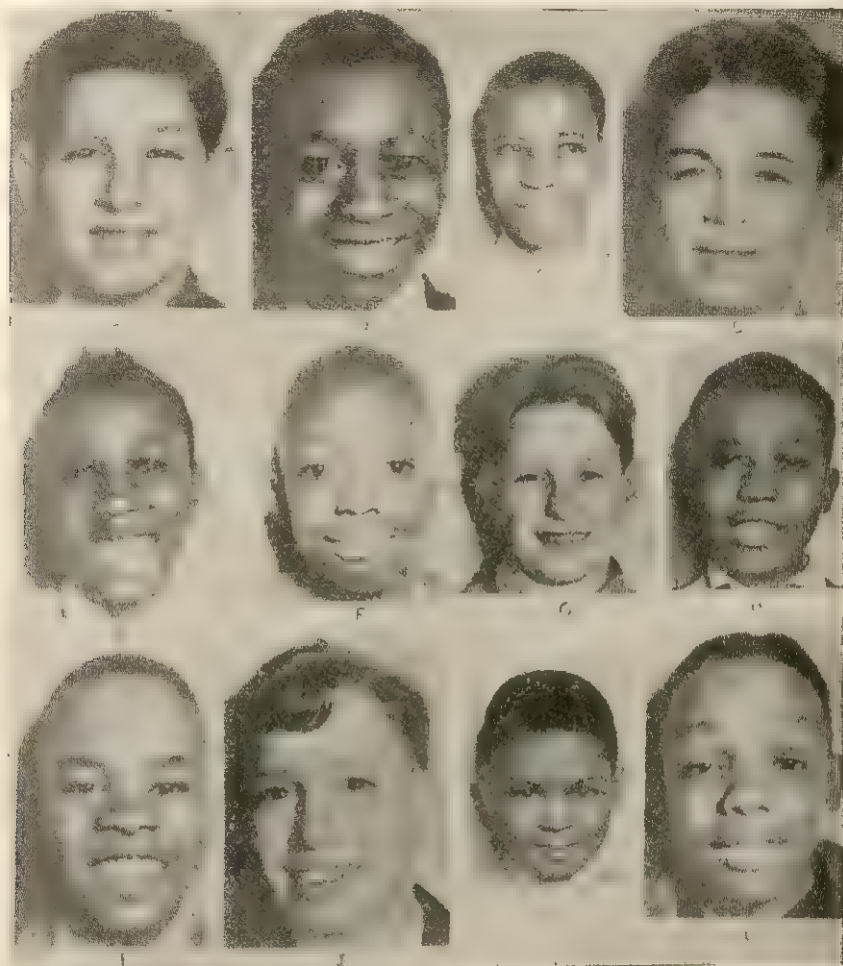


FIG. 36. Photographs used to measure attitudes toward Negroes. (By permission of Eugene L. Hartley and by arrangement with Harper & Brothers, New York.)

unequivocally either Negro or white. It will be noticed also that some of the Negro faces are lighter than the white faces while others are darker. (All pictures were of boys and were judged by boys.) The position of the white picture among the three Negroes is a chance one. The two tests administered to children from the kindergarten through grade 8 were to be answered directly from this picture, which in the original was about 10 by 10½ inches.

In the first test the subject was asked to "pick out the one you like best, next best, next best," etc., until all pictures were ranked. The scoring was directly dependent upon the ranks assigned the four white faces. For example, if they were ranked 1, 2, 3, 4, the score was 10, the lowest possible. Chance scoring of the ranks of white faces was 26.

In the second test, companions were selected from the pictures for a variety of imagined situations, as follows:

1. Show me all those that you want to sit next to you in a street car.
2. Show me all those that you want to be in your class at school.
3. Show me all those that you would play ball with.
4. Show me all those that you want to come to your party.
5. Show me all those that you want to be in your gang.
6. Show me all those that you want to go home with you to lunch.
7. Show me all those that you want to sit next to in the movies.
8. Show me all those that you would go swimming with.
9. Show me all those that you would like to have for a cousin.
10. Show me all those that you want to be captain of the ball team.
11. Show me all those that you want to live next door to you.
12. Show me all those that you like.

The same questions were used for both forms of the test.

In scoring this test the relative frequency was computed with which white faces were selected for all activities.

The third was the social-situations test. In its final form this test consisted of pictures in which there were children engaged in various activities in paired groups: (1) all white children, and (2) white children with one or more Negro children. The activities involved may be illustrated by playing baseball, dating in the ice-cream parlor, at home eating dinner, or in the workshop. The question asked the children whether they wanted "to join in with them and do what they are doing along with them."

The reliability of these tests can be only dimly perceived from the scores obtained from the same children after a period of 6 months. "Not only were group averages going up regularly, but relative position of children within groups was being maintained."

The validity of these tests was deduced more from their internal consistency and interrelation than with any measurement against outside measures of race attitude. Of the tests themselves, Test I, the ranks test, was more sensitive to race prejudice than either Test II or Test III. Test III was least sensitive. Some of the results indicate the tremendous possibilities in work of this kind. Prejudice appears at an early age, even in the kindergarten. There is some increase with the grades up to grade 4 or 5. Prejudice occurs about as often in New York as in Georgia or Tennessee, in urban as in rural areas. It appeared in mixed schools as

well as when the races were separated. Only children of Communists showed no prejudice toward the Negro.

E. C. Hunter's Test of Social Attitudes¹ contains opportunities for expressing one's attitude toward a single race, war, economics and labor, social life and convention, government, and religion. Before *each* statement there are five numbers —2, 1, 0, —1, —2—by which a subject can express five degrees of conviction from 2 if he is strongly convinced the statement is true, through 0 if he is undecided, to —2 if he is strongly convinced the statement is false. Norms are available at the college level only.

The question of race attitudes among children has been investigated.² There were no neatly scaled questions running from extreme favor to extreme disfavor but rather a set of natural situations about which questions were asked. Two samples follow:

- I. A Jewish family in Chicago was planning to buy a house and to move on a street where only native-born American families lived. The people who already lived on this street did not like to have this family move there, and went to the proprietor of the house trying to persuade him not to sell it to the Jews.
 1. Was it all right for the Jewish family to desire to move on this street?
 2. Was it all right for the people who already lived on the street to try to keep them from doing so?
 3. Ought a city to be divided into sections or quarters and each racial group to live in its own quarter?
 4. If the Jews had moved on the street, ought the other families to be friendly to them?
 5. Would Jews usually make as good neighbors as other people?
- II. A football team from Gordon College, Indiana, was to meet the team from Corliss College, Alabama. A Negro played on the Gordon College team. The team from the South sent word it would not play if the Negro was in the line-up. So the Gordon coach kept him out of the game.
 1. Did the coach do right in keeping the Negro out of the line-up?
 2. Would it have been better for Gordon College to cancel the game?
 3. Would it make any difference in deciding whether the Negro was to play if the game had been played in the South instead of the North?
 4. If you were a college athlete, would you just as soon play on a team with Negroes?
 5. If you were a college athlete, would you just as soon play against a team that had 2 or 3 Negro players as one that did not?

Questions could be answered in five ways, with scores as follows when "Yes, certainly" was considered the desirable answer:

¹ Hunter, E. C., *A Test of Social Attitudes*. Psychological Corporation, New York, 1936.

² Minard, Ralph D., *Race Attitudes of Iowa Children*, Studies in Character, Vol. 4, No. 2, University of Iowa, 1931. Items by permission of University of Iowa.

| | |
|------------------------|----|
| Yes, certainly..... | +3 |
| Probably yes..... | +2 |
| Uncertain..... | 0 |
| Probably no..... | -2 |
| No, certainly not..... | -3 |

If "No, certainly not" was considered the desirable answer the plus and minus signs were reversed.

Such situations dealing with attitudes toward Jews, Negroes, Filipinos, Chinese, Japanese, Mexicans, Italians, and foreigners in general were constructed. They were submitted to more than a thousand children and adolescents living in areas representing rural, semiurban, and urban populations.

The scores achieved could be compared with the opinions of those who might be considered authorities on race problems. Their judgments constituted what might be called the desirable race attitude. Only those questions were used on which 80 per cent of the authorities agreed. Some of the results indicate the value and possibilities of such a procedure.

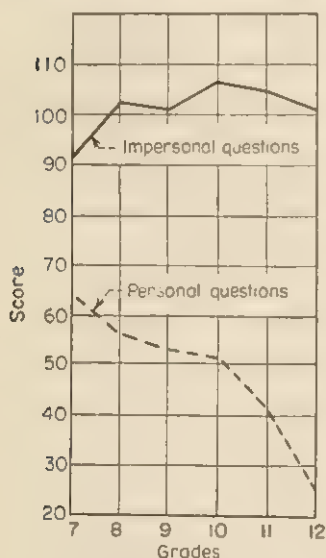


FIG. 37. Race attitudes as reflected by personal and impersonal questions (Minard, 1931). (By permission from The University of Iowa.)

There was little difference in the attitudes of boys and girls although the scores slightly favored the girls. A clear-cut difference did appear between the lower and the upper grades (tests were given from grades 7 to 12). There is a growth in race tolerance of an intellectual sort until grade 10, then perhaps a slight decline. On the questions which involve strong personal feeling there was a consistent

deterioration of attitude from grade 7 to grade 12 (Fig. 37).

Children's attitudes in general were far below the racial tolerance exhibited by the experts. Intelligence and race tolerance were slightly correlated, but there was no correlation with socioeconomic status.

THE USES OF AVAILABLE TESTS IN STUDYING POSSIBLE CHANGES IN ATTITUDES

It is evident that once attitude scales are carefully constructed they can be used to measure the influence of any activity on the attitudes being studied. Suppose, let us say, one wished to determine the influence

of warlike movies on the attitude toward war of children in the sixth grade. It would not be difficult to get them to express their attitudes toward war on Thurstone's scale before and after they had seen the movie. The difference between their first and second expressions of attitude would indicate the influence of the picture on their war attitude. Such an experiment has been made by two investigators (Thurstone and Peterson, 1933) who used the motion picture *All Quiet on the Western Front*. The showing of this picture did produce a demonstrable change on the attitude scale for war. Moreover, two such pictures produced a greater effect than one did. The change seemed to have lasted over a period of a year, but with some diminution in its amount.

In the other illustration from the same investigators it was shown that one motion picture changed the children's attitude toward the Chinese race by a measurable amount. The film *Son of the Gods*, which portrayed the Chinese in a favorable light, was shown to high school students. Before the showing of the film these young people had registered on a carefully prepared scale their attitudes toward the Chinese. The influence of the picture was definite and clear. The average attitude changed 1.22 steps on the scale. The change was in the favorable direction. This statistically reliable change had not disappeared in a year's time.

Further illustrations of studies in changing attitudes will now be introduced. In one study (Gardner, 1935) high school students were the subjects. These students after registering their attitudes were divided into three groups on the basis of their attitudes toward war. Three equivalent groups were formed. Group I listened to a carefully prepared lecture which appealed both to reason and emotion and which glorified war. After the lecture two stories were read which also glorified war. The other side of war was then presented; its cost in money, in men, and in materials.

In Group II the arguments in favor of war were presented but with no opposing ideas. Group III was the control. All students took the attitude scales at the end of the experiment. In Groups I and III the changes in attitude were negligible, but the students in Group II, those who for a few weeks had listened to lectures and heard stories glorifying war, changed their registered attitudes definitely toward war. Other studies have shown that students in colleges change their attitudes very little during their 4 years in college under ordinary conditions but that if specific attempts are made clear-cut changes are made. The reading of a single article on capital punishment, the reading of material justifying the Japanese invasion of Manchuria, and hearing President Roosevelt's speech on the Supreme Court changed the attitudes of the participants. Other studies have concluded that direct teaching on a

controversial issue will modify attitudes substantially. Without specialized teaching, changes in attitudes are slight.

Thus far this chapter has emphasized the importance of attitudes, the need of defining and describing them so that their presence can be detected, and the need of constructing scales on which changes of attitudes can be reflected. It has tried to make clear that attitudes are of tremendous importance in their effect on personality. So important are they that their development cannot be left to chance. Some illustrations have been offered of how attitudes may be modified. It was clear that they must be attacked directly and specifically, not indirectly and generally.

It is therefore recommended:

1. That we begin work immediately on deciding upon a small list of outstanding attitudes; those which most people would count desirable.
2. That we so define these attitudes and describe them that their presence can be detected in the behavior of young people.
3. That we then proceed to construct scales on which these attitudes may be reflected and apply them in such a manner that we can be very certain that our procedures are producing in attitudes the changes we want.

SUMMARY

Attitudes affect action. They are formed through (1) intense experience, (2) generalizations from several experiences, (3) acceptance by the individual from his mores, and (4) a splitting off from an already formed attitude. It was clear that it was not worth while to attempt the measurement of all attitudes but only of those which were of greatest importance in effective living. Since no over-all list of measurable attitudes had been made, lists of those considered important by small numbers of judges were made and appropriate tests and scales introduced. Attitudes must not be so general that they lose their applicability to concrete situations or so specific that the mere construction of scales for them would be prohibitive. The Thurstone technique of attitude-scale construction was introduced as a sample of the best that we have. Generalized attitude scales developed by Remmers after the manner of Thurstone were shown to be useful. Tests on beliefs on social issues showed much promise of measuring some beliefs and attitudes of real significance at the high school level. Experiments with pictures to indicate attitudes toward certain races and descriptions of situations emotionally loaded indicate a probable line of testing for the future. Scales of civic beliefs, tests of public opinion, and a conservative-radical opinionaire were introduced. They measured attitudes only indirectly. Finally, some attention was given to the importance of changing attitudes. It

was clear that such changes require specific, not general, instruction. Evidence was offered that scales are useful instruments for registering the amount of change.

QUESTIONS AND EXERCISES

1. What is an attitude? Name four or five important attitudes.
2. How are they learned? Why are they so important?
3. What lists of attitudes are there which indicate the outcomes of education?
4. How would you go about testing such attitudes?
5. What would an ideal attitude scale be? Compare with Thurstone's. On what principle are Thurstone's statements scaled?
6. What is The Bogardus' Scale of Social Distance? Evaluate it.
7. How do the scales of Remmers differ from those of Thurstone? Evaluate (a) the former's construction and arrangement, and (b) its principle of construction.
8. What are chief characteristics of Smith and Tyler's Test on Beliefs and Social Issues?
9. How are Hartley's pictures used to show attitude toward race? What did he discover?
10. Describe Minard's study of race attitudes. What was discovered about the change with age of attitudes toward race?
11. What other instruments are there for measuring attitudes?
12. Describe an experiment for the study of the change of attitude toward an important institution or idea. Name the scale you would use and describe precisely the learning procedures.

BIBLIOGRAPHY

Books

BARE, T. H.: "The Measurement of Attitudes," in T. H. Briggs, *et al.*, *The Emotionalized Attitudes*. New York: Bureau of Publications, Teachers College, Columbia University, 1940.

BRUNER, HERBERT B., and ARTHUR V. LINDEN: *A Tentative Check List for Determining the Positions Held by Students on Forty Crucial World Problems*. New York: Bureau of Publications, Teachers College, Columbia University, 1935.

LENZ, THEODORE F.: *C-R Opinionnaire (Conservatism-Radicalism)*. St. Louis, Mo.: Character Research Institute, Washington University, 1935.

LEWERENZ, ALFRED S., and HARRY C. STEINMETZ: *Orientation Test, 1935 Revision, for High School and College*. Los Angeles: California School Book Depository, 1935.

MURPHY, GARDNER, and RENSIS LIKERT: *Public Opinion and the Indi-*

vidual. New York: Harper & Brothers, 1938.

NEWCOMB, T. M.: "Social Attitudes and Their Measurement," in Gardner Murphy, Lois B. Murphy, and T. M. Newcomb, *Experimental Social Psychology*, rev. ed. New York: Harper & Brothers, 1937.

PETERSON, RUTH C., and L. L. THURSTONE: *Motion Pictures and the Social Attitudes of Children*. New York: The Macmillan Company, 1933.

REMMERS, H. H., and N. L. GAGE: *Educational Measurement and Evaluation, Attitudes and Related Aspects*, Chap. XVII. New York: Harper & Brothers, 1943.

SMITH, EUGENE R., RALPH W. TYLER, *et al.*: *Appraising and Recording Student Progress*, "Evaluation of Social Attitudes," pp. 203-244. New York: Harper & Brothers, 1942.

SMITH, F. T.: *An Experiment in Modifying Attitudes towards the Negro*,

Contributions to Education, No. 887. New York: Bureau of Publications, Teachers College, Columbia University, 1943.

THURSTONE, L. L., and E. J. CHAVE: *The Measurement of Attitude*. Chicago: University of Chicago Press, 1929.

WRIGHTSTONE, J. W.: *Wrightstone Scale of Civic Beliefs*. Yonkers, N.Y.: World Book Company, 1938.

Articles

BOGARDUS, E. L.: "A Social Distance," *Sociological and Social Research* (1933) 17:265-271.

CAREY, STEPHEN M.: "Professed Attitudes and Actual Behavior," *Journal of Educational Psychology* (1937) 28:271-280.

GARDNER, IVA COX: "The Effect of a Group of Social Stimuli upon Attitudes," *Journal of Educational Psychology* (1935) 26:471-478.

HINCKLEY, E. D.: "The Influence of Individual Opinion on Construction of an Attitude Scale," *Journal of Social Psychology* (1932) 3:283-296.

HORNE, E. PORTER: "Socially Significant Attitude Objects," *Studies in Higher Education*, XXXI, *Bulletin of*

Purdue University (1936) 37:117-126.

HOROWITZ, E. L.: "The Development of Attitude toward the Negro," *Archives of Psychology* (1936) Vol. 28, No. 194.

KELLY, IDA B.: "The Construction and Validation of a Scale to Measure Attitude toward Any Institution," *Studies in Attitudes, Studies in Higher Education*, XXVI, *Bulletin of Purdue University* (1934) 35:18-36.

LIKERT, RENSIS: "A Technique for the Measurement of Attitudes," *Archives of Psychology* (1932) Vol. 22, No. 140.

MINARD, RALPH D.: "Race Attitudes of Iowa Children," *Studies in Character*, Vol. 4, No. 2, University of Iowa, 1931.

PETERS, F., and M. ROSANNA: "Children's Attitudes towards Law as Influenced by Pupil Self Government," *Studies in Attitude, Series II, Studies in Higher Education*, XXXI, *Bulletin of Purdue University* (1936) 31:15-26.

REMMERS, H. H.: "Generalized Attitude Scales Studies in Social-Psychological Measurements," *Studies in Attitudes—A Contribution to Social-Psychological Research Methods, Studies in Higher Education*, XXVI, *Bulletin of Purdue University* (1934) 35:7-17.

CHAPTER 18

Measurement of Personality Traits

The term "personality" as used in this connection is a much narrower one than when used ordinarily. In its broadest sense personality may be thought of "as the total quality of an individual's behavior."¹ In this sense every instrument that has been studied in this text would be included as well as those contained in this chapter. But after instruments of measurement of achievement, intelligence, interests, and special capacities had been developed there arose a felt need for getting some sort of measurement of those other personality characteristics which loom so large both in the individuals adjustment and in his interaction with others. Personality inventories and tests came to include those traits and characteristics not included in tests of intelligence, of achievement, or of special capacities. More particularly, these inventories came to refer to those aspects of emotional adjustment which contributed to personality balance and integration. Many of these traits exhibit their leading characteristics when individuals' strongest desires are thwarted and they can find no satisfactory solutions for the resulting problems.

As in other areas of behavior, approach to a better understanding obtains through both the objective and the subjective techniques. Objectively, observation of behavior is made either systematically or fortuitously and then ratings are made of the behavior. Thus the experimenter takes his place at the front of a classroom and to one side in order to observe the oral habits-spasms or tics - which appear. He has in this manner narrowed the field of observation and may repeat the observation so that reliable results can be secured. Both the ratings of traits and the continued recording of chance observations concerning a student or pupil are objective in nature and are subject to errors of interpretation. On the other hand, if we can gain the cooperation of the subject himself so that he is willing to answer directly those questions which refer to his emotional life we can gain a great deal of information about his adjustment in a short time and with far less expenditure of energy. It is this subjective approach to the understanding and evalua-

¹ Woodworth, R. S., *Psychology*, 1940 edition. New York: Henry Holt and Company, Inc., p. 137.

tion of adjustment which first made its appearance and which has had far more research done upon it than the objective approach.

SELF INVENTORIES OR QUESTIONNAIRES

It was Professor Woodworth of Columbia University who in 1917 was called upon to develop a general screening test for our armed forces which would indicate the presence of the more severe types of mental maladjustment so that such cases could either be eliminated from the army program or made subject to further psychiatric investigation. He carefully collected symptoms of mental maladjustment which he edited, organized, and telescoped into the Woodworth Psychoneurotic Inventory.¹ Samples of this inventory are:

| | | |
|--|-----|----|
| 9. Does your heart ever thump in your ears so that you cannot sleep? | Yes | No |
| 19. Have you ever had fits of dizziness? | Yes | No |
| 29. Have you ever lost your memory for a time? | Yes | No |
| 39. Did the teachers in school generally treat you right? | Yes | No |
| 59. Do you ever have a queer feeling as if you were not your old self? | Yes | No |
| 79. Do you feel like jumping off when you are on high places? | Yes | No |

Such an instrument could be checked rather easily for internal consistency and reliability but its validity has been very difficult to determine.

FUNDAMENTAL DIFFICULTIES WITH SELF-INVENTORIES OR SELF-REPORTS

One of the most immediately perceived difficulties lies in the language itself. Take, for example, No. 9 above, which asks about the thumping of the subject's heart so that he cannot sleep. Is it to be interpreted as an emotional response in which the subject experienced fear accompanied by intense beatings of the heart so that he could not sleep? If this is all that is meant, then the experience is universal and has no symptomatic value. The investigator may have one thing in mind which the subject, though honestly trying, misinterprets. In the second place, the subject may not *know* the answer because what is called for has appeared and has been forgotten or repressed so that he is no longer aware of its presence. Thus, "Have you ever had fits of dizziness?" may be answered "No" when there has actually been a history of such spells in the patient's case and they have been forgotten. In the third place, there is the difficulty of getting the subject honestly to divulge the intimate experiences of his daily life. He may not want anyone to know that he has fits of dizziness or that he has lost his memory, and hence he may give the wrong answer. In the fourth place, it is very difficult to dis-

¹ Items by permission of C. H. Stoelting Co., Chicago.

cover personality traits that are really disparate segments of personality which do not overlap too much with other traits. Some of the dimensions studied are dominance-submission, introversion-extroversion, self confidence, self-sufficiency, etc. The burning question here is whether or not these dimensions are realities in the life of individuals or merely constructs in the mind of the tester. Are they independent of each other, or are they so closely related that one of them correlates highly with the other?

It is for these reasons that scores from psychoneurotic inventories are not to be interpreted as are simple reading scores or even those from intelligence tests. All scores must be regarded as tentative and experimental, as aids in interpreting the whole individual. For example, no one would be justified in interpreting a high score on the Bernreuter neurotic score as indicating a definitely neurotic condition. One could certainly interpret it as indicating that such a case needs further study or as confirmatory evidence of a condition which had already been suspected because of the activities of the subject. Such a score then must be interpreted in the light of the whole individual and not as a discrete entity. In general, three limitations of these inventories must ever be kept in mind. In the first place, in no case has the validity of an inventory been adequately determined. In the second place, the reliability of separate traits measured by the inventories are rarely high enough for individual diagnosis. One must remember that an individual's diagnosis based on a reliability as high as .90 is subject to considerable chances of error, the efficiency being 56 per cent. When profiles are constructed based on separate traits whose reliabilities are around .75, the consequences are almost ludicrous, for the efficiency of prediction is only 34 per cent. In the third place, the dimensions are not independent. A score on one dimension of personality may be made up partly of a score on a previous dimension. In a few cases, inventories have measured traits that are independent. Such independence involves rather elaborate statistical treatment called factor analysis. Even when the factors are computed all the investigator knows is that there is a Factor I, a Factor II, and Factor III which are uncorrelated, *i.e.*, are independent. The name which he gives to each factor is dependent upon its relations to various measures and his own psychological insight. In the first case, we have well-known psychological traits whose *independence is problematical*; in the second, we have definitely independent factors whose *names are problematical*. To sum up, the measurement and interpretation of personality traits are of great importance. The techniques of their measurement have not developed to the desired stage of validity. The interpretation based on scores from these instruments must be tentative and contingent. If proper precautions are exercised such instruments

are of very great importance in studying the personality of children and adults.

TYPES OF SELF INVENTORIES

Among the inventories for the study of adjustment at the senior high school and college levels, none has been more used than the Bernreuter Personality Inventory.¹ This instrument, composed of 125 items of the "yes—no—?" type, has the advantage of yielding six dimensions of personality after one administration. These dimensions are neuroticism, self-sufficiency, extroversion-introversion, dominance-submission, lack of self-confidence, and sociability. When the test was first constructed only the first four dimensions were used, but later it was learned that a rather high degree of relationship existed between these supposedly discrete dimensions. The coefficients of correlation between neuroticism and introversion, for example, was .95; between neuroticism and ascendancy, .81; and between neuroticism and self-sufficiency, .35. These high coefficients except in the case of self-sufficiency suggested to Flanagan² that a factor analysis might be made of these coefficients. The results of his work indicated that all the relations present were adequately covered by two factors: (1) lack of self-confidence, and (2) sociability. These two uncorrelated factors, Flanagan concluded, could be used for the entire four, and of the two, the lack of self-confidence was much more heavily weighted or, in other words, was much more important. Consequently two sets of scoring keys were constructed so that now the inventory may be scored in six different ways. Logically the first four dimensions should have been discarded and only the last two independent ones retained.

Some further facts about this inventory may be derived from a consideration of its construction. Four separate inventories went into the making of this test: Thurstone's Personality Schedule, Bernreuter's Self-sufficiency Scale, Laird's C Introversion Test, and Allport's A-S Test or test of ascendancy-submission. The items and problems of this test were ingeniously combined into the 125 items of the test and then scored in such a way that each dimension thereby derived would correlate highly with scores from the original test. For example, the dimension of neuroticism would have a very high correlation with Thurstone's Personality Schedule from which it was derived.

Samples taken from the Bernreuter Personality Inventory and their scoring on six different keys will be presented:³

¹ Stanford University Press. Items by permission.

² Flanagan, J. C., *Factor Analysis in the Study of Personality*. Stanford University, Calif.: Stanford University Press, 1935.

³ Items by permission of Stanford University Press.

1. Yes No ? Does it make you uncomfortable to be "different" or unconventional?
2. Yes No ? Do you daydream frequently?
3. Yes No ? Do you usually work things out for yourself rather than get someone to show you?
4. Yes No ? Have you ever crossed the street to avoid meeting some person?
5. Yes No ? Can you stand criticism without feeling hurt?

The scoring on the different keys for the first two items is shown in the accompanying table. In this manner the 125 items may be scored in six

| | Neurotic | Self-sufficiency | Extroversion-introversion | Dominance-submission | Lack of self-confidence | Sociability |
|--------|----------|------------------|---------------------------|----------------------|-------------------------|-------------|
| 1. Yes | 2 | -4 | 1 | -3 | 1 | -2 |
| No | -2 | 4 | -1 | 3 | -2 | 3 |
| ? | 0 | 1 | -1 | -1 | 3 | -3 |
| 2. Yes | 5 | 1 | 3 | -1 | 3 | 2 |
| No | -4 | -1 | -4 | 1 | -5 | -3 |
| ? | -2 | -2 | 0 | 2 | 0 | 5 |

different ways yielding thereby six dimensions of personality. Because each dimension is derived from all 125 items, its reliability is higher than if it were derived from 20 to 25 items as would have been the case had separate sets of items been responsible for the score in each dimension. The reliability coefficients for the first four dimensions range from .89 to .92 in one case and from .85 to .88 in the other; while the self-confidence dimension had a reliability of .86 and that of sociability, .78. "Such correlations would rate students 70% of the time on a five-step scale and practically all the time with an error of one step on each scale."¹

Since there are no readily available criteria against which to measure the results of this inventory the *validity* is necessarily in doubt. Bernreuter himself unfortunately presents in his manual coefficients of correlation between the dimensions secured from the scores of his test and the original tests from which his test was constructed as evidence of validity. This seems a trifle like begging the question, since his own personality inventory had in it many items from these very tests (*i.e.*, items from Thurstone's Personality Schedule, Laird's introversion-extroversion, etc.). For example, the r between Thurstone's neurotic

¹ Flanagan, J. C., "Technical Aspects of Multi-trait Tests," *Journal of Educational Psychology* (1935) 26:641-651.

inventory and Bernreuter's dimension of "neurotic" was .94, but Thurstone's neurotic inventory was the source of many of Bernreuter's items. One could hardly have expected a different outcome.

A more severe test of the *validity* of the Bernreuter Personality Inventory was made when it was used to test the emotional differences between the sane and the insane.¹ Three inventories by Woodworth, Bernreuter, and Page were found *not to be useful for distinguishing the normal from the insane*. Further evidence of evaluation appears in the contradictions in the results of the uses to which the inventory has been put. Two investigations found no great assistance from the inventory in differentiating between problem and nonproblem groups.² In one of these,³ correlations were made between the inventory scores and counselors' ratings, with very low coefficients as results. But problem cases frequently involve moral as well as emotional maladjustments and moral traits lie outside the boundaries of this inventory. On the other hand, another investigation⁴ found the inventory extremely valuable in distinguishing between well-adjusted and maladjusted students in a situation involving consultation service. It is possible that students came more nearly giving forthright responses when they knew that the scores were to be used to keep them adjusted more adequately. Likewise the Bernreuter Inventory was found to be of considerable value "as an aid in the diagnosis of psychopathic inferiors."⁵

It should be pointed out that Bernreuter's Inventory was not constructed to distinguish between the normal and the psychotic (insane) but between good and poor adjustment in otherwise normal subjects.

In summary, the Bernreuter Personality Inventory has been published for long enough time to discover something of what it can do. More than 100 studies have been made with its dimensions as the leading variables. The results are not clear. In one case, the self-confidence score was fairly valid but much less so was the sociability score. The inventory seems to be of more value in differentiating the emotionally maladjusted than in differentiating the psychotic. Criticisms generally leveled at self-inventories apply to this inventory. The subject has

¹ Landis, *et al.*, "Empirical Evaluation of Three Personality Adjustment Inventories," *Journal of Educational Psychology* (1935) 26:321-330.

² Jarvie, L. L., and A. A. Johns, "Does the Bernreuter Inventory Contribute to Counseling?" *Educational Research Bulletin* (1938) 17:7-9.

³ Speer, G. S., "The Use of the Bernreuter Personality Inventory as an Aid in the Prediction of Behavior," *Journal of Juvenile Research* (1936) 20:65-69.

⁴ Stogdill, Emily, and Minnie E. Thomas, "The Bernreuter Personality Inventory as a Measure of Student Adjustment," *Journal of Social Psychology* (1938) 9:299-315.

⁵ Hathway, S. R., "The Personality Inventory as an Aid in the Diagnosis of Psychopathic Inferiors," *Journal of Consulting Psychology* (1939) 3:112-117.

difficulty in interpreting the items such as: "Do you daydream frequently?" The subject may also lie about an item even if he understands it. On the positive side, this inventory does have two dimensions that are uncorrelated or independent. If, as Flanagan's study indicates, 78 per cent of the variance discovered is due to Factor F_{1-C}, which is the lack of self-confidence, then Bernreuter's Inventory might best be used for measuring this trait. Thus the lack of self-confidence looms up as one of the best measured of the dimensions of personality.

Another self-inventory resembling in general outline the one just described is the Bell Adjustment Inventory¹ which also has an adult form and a student form. This instrument was constructed from the 223 items of the Thurstone Personality Schedule and an additional 188 new items supplied by the author. As with other such tests, each item was checked against the items as a whole by discarding those which did not differentiate between the "upper and the lower 15 per cent of the scores in the distribution for each category."² In addition, the criterion of applicability (*i.e.*, the item must be checked by at least 25 per cent of the maladjusted group) was also applied for retention of an item as well as that one which eliminated the items which were sometimes misunderstood. From this rather rigorous process 140 items remained to compose the inventory. Four categories, under which there are 35 items each, are (1) home adjustment, (2) health adjustment, (3) social adjustment, and (4) emotional adjustment. The adult form adds another: occupational adjustment. The author claims that these divisions are concrete and objective and that the counselor and his counselee understand these terms. The overlapping among the divisions is not large. Intercorrelations among the four divisions range from .04 to .53 and average .35.

The reliability of the inventory, shown in the accompanying table, is as high as one could expect.

| Division | Reliability |
|---------------------------|-------------|
| Home adjustment..... | .89 |
| Health adjustment..... | .80 |
| Social adjustment..... | .89 |
| Emotional adjustment..... | .85 |
| Total score..... | .93 |

The inventory can be easily scored in a few minutes by simply counting the number of items marked in each division. It may also be scored with weights ranging from +6 through 0 to -6. These weighted scores are slightly more reliable (.95 as compared with .93 for the total score) and

¹ Stanford University Press.

² Bell, Hugh M., *The Theory and Practice of Personal Counseling*, page 25. Stanford University, Calif.: Stanford University Press, 1939.

correlate .96 with the unweighted scoring. *The author of the inventory recommends the use of the unweighted key.* The probable error of measurement, which provides the limits within which each true score may be found, has also been computed.

The author has provided a scheme, illustrated in the accompanying table, whereby scores can be changed into descriptive phrases. This

| | High school score range | | Descriptive | College score range | |
|-----------------|----------------------------|----------------|-------------------|------------------------|----------------|
| | Men (161) | Women (190) | | Men (171) | Women (243) |
| Home adjustment | 0-1 | 0-2 | Excellent | 0-1 | 0-1 |
| | 2-4 | 3-5 | Good | 2-4 | 2-4 |
| | 5-9 | 6-13 | Average | 5-9 | 5-9 |
| | 10-16 | 14-20 | Unsatisfactory | 10-16 | 10-15 |
| | Above 16 | Above 20 | Very satisfactory | Above 16 | Above 15 |

helps the counselor give a practical interpretation of the obtained score.

The validity of the inventory has been well attended to. In the first place, the item selection whereby only those items were chosen which differentiated between the upper and lower 15 per cent of individuals in a total distribution is, in itself, a validating procedure. In the second place, each section was evaluated by means of interviews with 400 college students extending over 2 years. In the third place, the inventory was correlated with those other inventories which purported to measure the same functions. For example, the inventory's measure of social adjustment was measured against Allport's test of ascendance-submission and Bernreuter's measure of dominance-submission, while the emotional-adjustment section was correlated with Thurstone's schedule. Then, too, the test as a whole was measured against Thurstone's Personality Schedule. These validating coefficients ranged from .58 to .89. Omitting the two coefficients with Thurstone's Schedule, which would be expected to be high in the light of the inventory's construction, the validating coefficients ranged from .58 to .79. In the fourth place, and best of all, the inventory was measured against the judgment of groups of students selected by counselors and school administrators. Two groups were formed by these counselors: the well-adjusted group and the poorly adjusted group. For example, in the area of the home 51 were well adjusted and 51 poorly adjusted; in health 42 were well adjusted and 42 poorly adjusted. In like manner two groups were formed in the social and emotional areas. The question to be answered was: "Does the inventory show a statistically reliable difference between the mean

of the poorly adjusted group and the mean of the well-adjusted group in each of the four sections of the test?" The answer was categorically yes. In every section the inventory distinguished between the well-adjusted group and the poorly adjusted group. Another way of looking at the differences between these two groups is by calculating the overlap between the scores received by the well-adjusted and poorly adjusted groups. The percentages of one group which reached or exceeded the median of the other was nowhere greater than 14 per cent, and in two cases as low as 2 per cent, whereas the percentages would have been 50 had the two groups scored the same. Item analysis was also made by comparing scores made by high school girls and college girls, high school girls and high school boys, and college girls and college boys. The successes and failures of delinquent boys and girls were also compared with each of the above groups.

The correlations between scores on intelligence tests and on Bell's Inventory and between this inventory and college scholarship are no larger than chance. Nor does the inventory differentiate between successful and unsuccessful teachers, although the unsuccessful teachers make slightly higher scores than the successful ones, thus indicating greater maladjustment.

Those who have used the Bell Inventory for counseling bear witness to its capacity to pick out a high percentage of adjustment difficulties which a careful clinician would find and to the fact that it misses only a small proportion of such difficulties. The most captious critics admit that the inventory has everything but validity. One must remember that these results were secured under the most favorable conditions. If there is complete rapport between the subject and the experimenter such results as have been described are possible.

The California Test of Personality¹ is really a set of inventories distributed as follows:

1. Primary A (kindergarten through grade 3)
2. Elementary B (grades 4 to 9)
3. Intermediate B (grades 7 to 10)
4. Secondary A (grades 9 to 14)
5. Adult series

All the members of the set are built on the principle that personality "refers rather to the manner and effectiveness with which the whole individual meets his personal and social problems, and indirectly the manner in which he impresses his fellows." In each inventory the scores received are divided into two main divisions:

1. Self-adjustment, under which are subsumed (a) self-reliance, (b) sense of personal worth, (c) sense of personal freedom, (d) feeling of

¹ California Test Bureau, Los Angeles, Calif.

belonging, (e) freedom from withdrawing tendencies, and (f) freedom from nervous symptoms.

2. Social adjustment, with its subdivisions of (a) social standards, (b) social skills, (c) freedom from antisocial tendencies, (d) family relations, (e) school relations, and (f) community relations.

Up through the inventories suitable for grade 10, these divisions are printed in plain view upon the inventory which the subject takes. In the secondary A and adult series only a few letters in each word are used so that their meaning would be unrecognizable. The reliabilities computed by the split-half method and then substituted in the Spearman-Brown formula are as shown in the accompanying table. These

| | | 1. Self- adjust- ment | 2 Social adjust- ment | Total com- ponents |
|-------------------------------|--------------------|--------------------------------|--------------------------------|--------------------------|
| Intercorrelation of 1 and 2 = | .66 Primary A | .893 | .873 | .922 |
| | .66 Elementary B | .888 | .867 | .933 |
| | .74 Intermediate B | .898 | .872 | .932 |
| | .54 Secondary A | .904 | .908 | .931 |
| | .76 Adult series | .888 | .898 | .918 |

coefficients of reliability are high enough for individual diagnosis for the inventory as a whole and for the two major divisions. It is certainly open to question as to whether or not scores on components of the inventory "are sufficiently high to locate more restricted areas of personality difficulty." From the table also it is clear that the correlations between the two large divisions of the inventory vary from .54 (marked or substantial) to .76 (high). Even the main divisions of this test are far from being uncorrelated.

The claims for validity of this set of inventories are based more largely upon the manner of their construction than on the results of their use. The items for the first four inventories are based upon a study "of over 1000 specific adjustment patterns or modes of response to specific situations which confront children of these ages. Many of these items had previously been validated by other workers." The bases for selecting the items for the final form of the tests were four:¹

- (a) The judgments of teachers and principals regarding their relative validity and significance. (b) The reactions of pupils expressing

¹ *Manual of Directions, California Test of Personality, Elementary Series*, p. 2. By permission from California Test Bureau, Los Angeles, Calif.

the extent to which they felt confident and willing to give correct responses. (c) A study of the extent to which pupil responses and teacher appraisals agreed. (d) A study of the relative significance of items by means of the bi-serial r technique.

This biserial r technique is a procedure by means of which each item is correlated with the total to measure its degree of agreement with the test as a whole. The manual furnishes only this bare outline of the selection of items without any statistical confirmation. There was some attempt to disguise the meaning of the items in the various inventories so that their intent would not be too apparent; e.g., the item is *not* "Do you cheat?" but "Are some people so unfair that you try to cheat?" and not such a question as "Are you mean to people?" but rather "Are people often so bad that you have to be mean to them?"

This low visibility of items, however, is immediately negated in the first three inventories by clearly printing the names of the components on the face of the inventory which the child takes.

The procedures for administering, instructions for scoring, and norms are all that could be desired. The norms are percentile scores easily read from the tables so that the profile consisting of the main divisions and the twelve subsections may be easily constructed. Were these total scores certainly valid and the subscores valid and reliable, and did they not overlap the one with the other, no more desirable graph could be constructed than this one based on such important outcomes of education. While the profiles are useful, they must be received and considered in the full light of how uncertain their meanings are. Above all, we must remember that these scores are based on what subjects *say* they feel. In conclusion, there is no doubt but that this series of inventories are as useful for school purposes as any others. The manual uses five full pages to treat of individuals who deviate too far from the normal. It is questionable whether the discussion of so difficult a problem in such a small space might not be a dangerous procedure.

Another inventory suitable for use with younger children (grades 4 to 9) is Aspects of Personality by Pintner, Loftus, Forlano, and Alster.¹ Three dimensions of personality, practically uncorrelated with each other, are obtained directly from the scores: (1) ascendance-submission, (2) extroversion-introversion, and (3) emotionality. The items for this test were selected from seven inventories which had already been constructed and also from new items which were developed.² The language

¹ World Book Company, Yonkers, N.Y. Items by permission.

² Woodworth-Matthews's Psychoneurotic Questionnaire; Allport's A-S Reaction Study; Thurstone's Personality Schedule; Pintner's General Opinion Test; Bernreuter's Personality Inventory; Maller's Character Sketches; Lecky's Individuality Record.

of the items was simplified in order to fit the level of fourth-grade children. When the children could not understand the items they themselves were allowed to reword them. From the inventories and from the new items suggested by the authors about 900 statements were secured. These items were rated by the authors on the bases of relevance and importance. Unanimous agreement of the authors as to the adequacy of each item was required for the preliminary tryout. In order for the dimensions of personality to be as independent as possible, provision for independence of measured traits was made in the construction of the tests. Each item must have a biserial r of at least .30, .40, or .50 for inclusions in the categories of ascendance-submission, extroversion-introversion, or emotionality. In the second place, items included in one section must not have a high correlation with the other two sections. As a matter of fact no item was included which correlated .14 or more with the score of either of the other two sections. Thus that principle of good construction, of having an item correlate high with the dimension under which it falls and low with the others, was carried through.

A rather interesting personalization of the items was secured as follows:

| | | |
|--|----------------------------|----------------------------|
| I don't like to ask questions in class | <input type="checkbox"/> S | <input type="checkbox"/> D |
| I like to play rough sports | <input type="checkbox"/> S | <input type="checkbox"/> D |
| I feel tired most of the time | <input type="checkbox"/> S | <input type="checkbox"/> D |

The subject was to think whether he was same (S) or different (D) and cross out the proper letter.

The reliability of the inventory was tested out both by the split-half and the test-retest procedures. For the Ascendance-Submission dimension the coefficients were .69 and .65; for the Extroversion-Introversion dimension, .70 to .76; and for the Emotionality dimension, .79 to .91. This would indicate reliability sufficiently high for use with groups of children but inadequate for individual analysis. Its best use probably would be in follow-up questions based on the reactions to the individual items. Percentile standards are furnished in the manual along with suggestions as to what to do with children who score very low in each of these dimensions. But the number of cases used in the development of the standards is not mentioned. The manual itself ends up with a general caution; it advises that "no simple group test of this type can diagnose, but it can indicate children who need careful attention." However, one must not regard percentile scores derived from group inventories as anything more than "a general description of a child's personality" and must not expect it to be a too accurate diagnosis of personality difficulties.

As in all other such inventories the validity is weak. It has not been measured against any sound criteria outside itself. Possibly also the instructions are a little infantile for children in grades 8 and 9. One has the distinct impression that it is pitched on a fourth- or fifth-grade level. There is also no definite suggestion in the manual for careful observations of behavior as supplementary to the inventory's scores.

The Maller Case Inventory¹ is constructed somewhat differently from the inventories already described. It is divided into four parts:

1. Controlled association test
2. Adjustment test
3. Honesty test (self-scoring)
4. Ethical judgment test

In the controlled association test, 50 words were selected out of a list of 200 items. These 200 words were collected from the Kent-Rosanoff, Jung, and other lists of free-association items. In these original lists a word is given and the subject responds with the first thing that comes to mind. It was found that there were "usual" responses for normal subjects and individual or unusual responses for subjects with some emotional maladjustment. Using this procedure Maller selected responses which (1) were usual, and (2) were "uncommon, personal, emotional, or involving superstitious ideas." He tried out these 200 items on (1) adult insane groups, (2) adult normal groups, (3) probationary children, and (4) normal children and found that 50 items distinguished clearly between the normal and the probationary children. Here are two sample items:

- 6 *Black* *death* *white* _____
 16 *Foot* *hand* *paralyzed* _____

The subject underlines one of the two words on the right which is connected with the key word on the left or he may write in a word.

The adjustment test is made up of items selected from Maller's own character sketches. The selection of items was based on a "thorough item analysis comparing the responses of well adjusted children and adults with those of serious problem cases, delinquents, and psychiatric patients. The undesirable responses involve extreme introversion, lack of self-control, feeling of inadequacy and inferiority, and symptoms of psychoneurotic tendencies." The subject is asked whether he feels the same (S) or different (D).

4. Sometimes has a feeling that things are not real.
14. Hates people who tell him frankly what they think about him.

¹ Teachers College, Columbia University. Items by permission.

Part 3, the honesty test, is composed of items selected from Maller's Test of Sports and Hobbies. Some of the items are so difficult that any claim as to their knowledge may be immediately suspected of dishonesty. For example, "The longest boxing match on record was held in 1893."

Probably the scores on honesty should not be interpreted as "honesty scores" but as "honesty-in-a-written test scores." In brief, the scores are specific, not general. Furthermore, the author's experience with the double-testing technique in studying deceit convinces him that many astute older children would catch on to the purpose of Section III, thereby causing the results to be untrustworthy.¹

In Part IV the items were selected from Maller's Ethical Judgment Test. Five items were selected on the basis of their discrimination between delinquents, probationers, etc., and normal groups. One sample is:

Philip dropped and broke a victrola record that was his mother's favorite. He knew that his mother would feel badly about it.

_____ Philip told his mother what had happened.

_____ Philip hid the record and didn't say anything about it.

The inventory also has a series of questions at the end inquiring about the subject's socioeconomic status, interests in books, recreation, and movies, and a series of items dealing with wishes, fears, and worries.

The reliability of the test is quite satisfactory, varying from .90 to .96 for the four separate divisions and from .93 to .94 for the inventory as a whole. The author states that the inventory has been used to compare some 400 problem cases with 5,000 normal ones and for purposes of studying delinquents. Its tentative norms are based on a population of 5,214 cases but are reported largely for the inventory as a whole and not for each of the four parts. This latter is most desirable.

The validity of this inventory checked in its construction and in its application is still a very weak feature. No attempt has been made to furnish the validity based on the stricter criterion of correlation. Such correlations would be desirable for each of the sections. The purpose of the test is not so clearly indicated as in some other inventories. The author really owes it to his public to indicate the tentative nature of the scores, which are to be used as leads rather than as personality facts. While this inventory shows promise, much needs to be done on the development of norms and of validity before it can be very useful.

¹ Jordan, A. M., "Cheating in the Classroom with Emphasis on the Influence of Friends," pp. 437-471 in Kelley, T. L., and A. C. Krey, *Tests and Measurements in the Social Sciences*. New York: Charles Scribner's Sons, 1934.

The Rogers Test of Personality Adjustment,¹ useful for ages 9 to 13, is unique on two counts. Its validity was assured by basing the norms "upon a study of fifty-two problem children," on whom an exhaustive study had been made. In the second place, the scores are looked upon merely as a numerical summary, as an opportunity for the detailed study of the individual case. The divisions used are based on four so-called diagnostic scores divided according to practical considerations.

1. The personal-inferiority score indicates the degree to which the child thinks himself to be physically or mentally inadequate—*i.e.*, duller, weaker, less good looking, less capable, than his companions.

2. The social-maladjustment score measures the extent to which he is unhappy in his group contacts, poor at making friends, poor in the social skills.

3. The family-adjustment score indicates whether jealousy of parents or siblings is present, whether there is a feeling of being unwanted, or whether there is too much dependence on one or both parents.

4. The daydreaming score indicates the extent of the child's fantasy life and, taken with the other three scores, shows how the child is solving his problems.

These four "diagnostic" scores are derived from six "tests," the first and second of which have to do with wishes. The first has to do with the sort of person you would wish to be could you change yourself; the second, with the fulfillment of your wishes. One might want to be a policeman, a singer, a lawyer, or a poet and one might wish to be stronger, to be bigger, to have more money, or to be better looking than at present. "Test" 3 is built on the old problem of writing down the names of three people you would choose to take with you if you were going away to live on a desert island.

"Test" 4 is best explained by an illustration.

Mary is the prettiest girl in school

Am I just like her?

yes

| | | | | | | | |
|--|--|--|--|--|--|--|--|
| | | | | | | | |
|--|--|--|--|--|--|--|--|

 no

Do I wish to be just like her?

yes

| | | | | | | | |
|--|--|--|--|--|--|--|--|
| | | | | | | | |
|--|--|--|--|--|--|--|--|

 no

Other samples may be had by substituting for the first part of the above illustration: "Gladys has the nicest clothes of anyone in school," "Lucile is a leader. The girls all do what she wants them to do," "Anna is the most popular girl in school. Everybody likes her."

In "Test" 5 an attempt is made to obtain the child's own estimate of his possession of certain important qualities. Two examples are:

¹ New York: Association Press. Items by permission.

How many friends would you like to have?

- a. none
- b. one or two
- c. a few good friends
- d. many friends
- e. hundreds of friends

Do people treat your brother better than they treat you?

- a. never
- b. sometimes
- c. often
- d. almost always
- e. I haven't any brother or sister.

In the sixth and last section of this personality inventory the subject is asked to list his parents and siblings, his best boy friend and best girl friend according to their ages and then "put a '1' in front of the person you love most, a '2' in front of the person you like next best and a '3' in front of the person you like next best."

The scoring of the test is rather complicated. However, detailed instructions for scoring are given which are not too difficult to follow. Norms are based upon the scores of 167 children. High scores are always indicators of maladjustment and ranges of scores are given which define the terms "low," "average," and "high" for each of the four divisions. Finally the manual gives four case histories and explains exactly how this inventory aids in their interpretation.

The reliability of the inventory is around .70, a score unacceptable for the diagnosis of the individual case but perhaps high enough to be used in the clinic where the whole case history had been made. At least one clinician says, "We have used this test in our clinics, and with the exception of the time-consuming method of scoring, have found it the most satisfactory instrument of personality measurement."¹ In spite of this encomium, the test lacks independence in its divisions. Furthermore, the very validities on which its claims to excellence lie are based on low correlations (.38 to .48) between the scores of the tests and clinicians' ratings. In both the divisions of family maladjustment and personal inferiority the correlations with ratings fall below .40. This inventory is a clinician's attempt to bring objectivity and measurement to the support of personal observation and rating. As such it should be commended, but still we should not be satisfied with a reliability of .70, a "natural" division into four dimensions, and a too-complicated scoring procedure.

¹ C. M. Loutit, *Nineteen Forty Mental Measurements Yearbook* (Oscar K. Buros, ed.), Item 1258. Highland Park, N.J.: The Mental Measurements Yearbook, 1941.

TABLE 17. LIST OF PERSONALITY INVENTORIES NOT INCLUDED IN TEXT

| Name | Grade | Validity and contents | Reliability | Publisher |
|---|---|--|--|----------------------------------|
| Bell School Inventory | High school | Items selected to differentiate between upper and lower 15 per cent of 450 high school students. Measures adjustments to (1) fellow students, (2) school plant, (3) school organization and offerings, (4) school administration, (5) teachers | .94 | Stanford University Press |
| Link Inventory of Activities and Interests | 7-13 | Part 1 checks interests in games and studies. Part 2 is 150 items from which scores can be obtained: (1) personality, (2) social initiative, (3) self-determination, (4) economic self-determination, (5) adjustment to opposite sex. May derive P.Q. (personality quotient) | For parts, .78-.88 | Psychological Corporation |
| Brown Personality Inventory for Children | 4-9 | 80 items. Total scores analyzed into (1) home, (2) school, (3) physical symptoms, (4) insecurity, (5) irritability. Questions are open and evident. Validity based on literature about the neurotic child | .90 | Psychological Corporation |
| The Detroit Adjustment Inventory (H. J. Baker) | Junior and senior high school | 120 items assembled around 24 topics including health, physical status, worries, fears, anger, pity, introversion, home status, reactions to school, sportsmanship, and morals. Depends largely on clinical evidence for validity | | Public School Publishing Company |
| Minnesota Personality Scale (John G. Darley and Walter J. McNamara) | College and last 2 years of high school | Fits best at college level. Measures (1) morale, (2) social adjustment, (3) family relations, (4) emotionality, (5) economic conservatism. Validity by measuring scale scores against clinically diagnosed maladjustments. Low r between parts | .84-.97 for parts | Psychological Corporation |
| Loofburrow-Keys Personal Index | 7-9 | Items selected which distinguish significantly between serious disciplinary problems in junior high school and an unselected group. Four tests: (1) false vocabulary, (2) social attitudes, (3) virtues, and (4) adjustment questionnaire. In follow-up study with same population "74 per cent (of changes) were in the direction indicated by the Personal Index (manual, p. 7)" | .84-.92 for the divisions; .95 for the whole inventory | Educational Test Bureau |

In Table 17 there are described personality inventories not included in the body of the text.

THE VALIDITY OF PERSONALITY INVENTORIES

Indications of the validity of personality inventories have been introduced throughout this chapter. In the present discussion, evidence will be brought forward to clinch the idea that *such instruments must be used with the greatest care*. This evidence has been collected and evaluated by Ellis.¹ He quoted directly from the investigators who had used these

TABLE 18. VALIDATION OF PERSONALITY INVENTORIES—NEUROTICISM
OR INTROVERSION*
(STUDIES OF DIFFERING TYPES. ELLIS, 1946)

| | Number | Positive | Questionably or mainly positive | Negative |
|--|--------|----------|---------------------------------------|----------|
| 1. Total | | | | |
| 1. By behavior problem. Diagnosis. Subjects mainly children.. | 9 | 2 | 1 | 6 |
| 2. By diagnosis of delinquency... | 34 | 15 | 6 | 13 |
| 3. By psychiatric and psychological diagnosis..... | 75 | 36 | 9 | 30 |
| 4. By rating diagnosis. Ratings by teachers, friends, or associates..... | 44 | 12 | 10 | 22 |
| Total..... | 162 | 65 | 26 | 71 |
| 2. By inventories | | | | |
| Bell adjustment inventory..... | 12 | 1 | 0 | 11 |
| Bernreuter Personality Inventory.. | 29 | 9 | 6 | 15 |
| Thurstone Personality Schedules.. | 10 | 4 | 1 | 5 |
| Woodworth Personal Data Sheet.. | 29 | 11 | 4 | 14 |
| Other personality tests..... | 82 | 40 | 15 | 27 |
| Total..... | 162 | 65 | 26 | 71 |

* By permission of *Psychological Bulletin*.

inventories, and then summarized and quantified the results. His treatment covers for the most part those studies which utilized inventories claiming to test neuroticism or introversion. Only objective results are considered. In Table 18 is indicated the degree of success which the inventories achieved. In this table, four lines of evidence are introduced

¹ Ellis, Albert, "The Validity of Personality Questionnaires," *Psychological Bulletin* (1946) 43:385-440.

which were derived from the actual application of inventories to real life situations. In behavior-problem diagnosis, inventories were administered to groups of behavior-problem children and their results compared with those secured from normal groups. In diagnosing delinquency, test results of delinquents are compared with those of normal groups. In psychiatric and psychological diagnosis, results from case studies by psychiatrists or psychologists are compared with results from the inventories. In *rating diagnosis*, teachers, friends, or associates rate individuals and these ratings are compared with scores on the inventory. By inspecting the totals in the table it is evident that out of 162 studies only in 65 cases did the inventories clearly differentiate between the groups studied. The author concludes (page 426),

It is concluded that group-administered paper and pencil personality questionnaires are of dubious value in distinguishing between groups of adjusted and maladjusted individuals, and that they are of much less value in the diagnosis of individual adjustment or personality traits.

While the author of this text does not subscribe entirely to such devastating criticisms of personality inventories, he realizes the importance of extreme care in the inferences drawn from the results of the administration of personality inventories to children and high school students.

RATING SCALES

Another procedure used for securing quantitative expressions of personality traits is that of rating. Aspects of rating have already been apparent in the neurotic inventories, in the measurement of interests and in the expressions of attitudes. But in each of these three procedures the rating was largely self-rating. When an individual answers such a question as "Do you feel miserable most of the time?" by marking "Yes," "No," or "?" he is rating himself upon the possession of this trait. Again, when an individual is disclosing his attitude by indicating his degree of belief in a statement he is rating himself. For example, when he indicates his belief in the statement, "Segregation of Negroes in trains, restaurants, theaters, hotels, and schools should be required by law" (voting 2, 1, 0, -1, or -2, in which 2 indicates strong conviction that the statement is true, and -2 an equally strong conviction that the statement is false, with the other numbers indicating intermediate positions of belief), he is rating himself on an attitude scale.¹ In like manner, when an individual expresses his liking for,

¹ Hunter, *op. cit.*

indifference of, or dislike of public speaking, musical comedy, or making a radio set¹ he is rating himself in interest. The conditions under which self-rating is worth while have already been described.

The present problem is none of these. It has to do with the *registering in a quantitative way of the presence of certain traits in individuals by observers*. Here is clearly the difference between introspection and observation. In the previous paragraph have been described the results of introspection. In the present, the reader is being introduced to the results of observation. In the rating of others, errors are apt to arise because (1) the observer himself has his own biases and Bacon's idols by means of which the behavior of others is colored in its interpretation, (2) the observer has not seen enough instances of the trait in question to be able to say that the individual possesses a great amount of this trait or not. Thus the rater may jump to conclusions (the inductive leap) from one or two observations. For example, the author saw a retarded 13-year-old boy hurl a regulation baseball into the midst of a group of smaller boys whose baseball it was by rights. This act hurt seriously one small boy. If you then rated such a boy on the trait of cruelty would you rate him high or not? The rating of others, however, has one fundamental advantage over self-rating: it need not be affected by self-interest. Whereas the impression which a self-rater gives of himself may be colored by his own fears of placing himself in an embarrassing position, the rating of others need have none of this at all.

The question of the separateness or independence of personality traits also plagues the rater. When one trait has been rated, can another trait be found whose rating can take place without being disturbed by the first rating? Unless relatively independent traits can be found, the correlation between them will be high (the "halo" error), and it will be impossible to discover whether they are really related or merely have a high correlation because of the influence of the rater's attitude.

TYPES OF RATING SCALES

The forms which these scales take are intended to improve the accuracy and ease of rating and to help the rater put his ideas in a quantitative form.

In the first form a line, usually about five inches long, is used. It is divided into five or more equal divisions and underneath each linear division is placed a verbal description of that amount of the trait possessed. A good illustration of this type is Schedule B of the Haggerty-Olson-Wickman Behavior Rating Schedules.²

¹ Strong, E. K., *Vocational Interest Blank for Men*. New York: Psychological Corporation, or Stanford University, Calif.: Stanford University Press.

² By permission of World Book Company, Yonkers, N.Y.

6. *Is he mentally lazy or active?*

| | | | | |
|------------------------------------|---------------------------------|----------------------------------|--------------|---------------------------------|
| Interests lazy and inert (5) | Lethargic idles along (3) | Is ordin- arily active (2) | Eager (1) | Shows hyper- activity (4) |
|------------------------------------|---------------------------------|----------------------------------|--------------|---------------------------------|

11. *What is his physical output of energy?*

| | | | | |
|------------------------------|--------------------------|-------------------------------------|-------------------------------|---|
| Extremely sluggish (5) | Slow in action (3) | Moves with required speed (2) | Energetic Vivacious (1) | Overactive Hyperkinetic Meddling (4) |
|------------------------------|--------------------------|-------------------------------------|-------------------------------|---|

24. *What tendency has he to criticise others?*

| | | | | |
|----------------------------|-----------------------------|--|--------------------------------------|---|
| Never criticises (3) | Rarely criticises (1) | Comments on outstanding weaknesses or faults (2) | Has a critical attitude (4) | Extremely critical, rarely approves (5) |
|----------------------------|-----------------------------|--|--------------------------------------|---|

One's rating is indicated "by placing a cross (X) immediately above the most appropriate descriptive phrase." The numbers at the bottom are used for scoring and are derived from the behavior scores which the individuals so rated received. For example, in No. 11 the children rated as extremely sluggish "had an average behavior score of 44.9 on Schedule A" while the overactive ones averaged 27.1. For this reason the first descriptive phrase in No. 11 is rated as 5 and the last one as 4. To secure a behavior score for an individual, simply add up the scores of the scales on which he is rated. The larger the score, the greater the number of personality difficulties.

There are many variations of this type (of rating scale). In some scales the line is continuous; the points are defined, but one is permitted to make intermediate judgments by checking at any point along the line. Another variation is simply to have the line and numbers at the division lines instead of descriptive phrases or else the same phrases at each division in every trait. For example:

| | | | | |
|-----------|--------|----------|--------|------------|
| Extremely | Rather | Somewhat | Hardly | Not at all |
|-----------|--------|----------|--------|------------|

Still another attempt at more exact definition of the trait is as follows:¹

¹ Filer, H. A., and L. J. O'Rourke, "Progress in Civil Service Tests," *Journal of Personnel Research* (1923) 1:484-520. Items by permission from *Personnel Journal*.

| | | | | | |
|---|-------------------------------------|---------------------------------------|-----------------------------|------------------------------------|--|
| Attitude toward work: Consider voluntary interest and effort in work | Unconcerned and no voluntary effort | Interest and effort below average | Average interest and effort | Interest and effort above average | Shows keen interest and whole hearted effort |
| Neatness: Consider orderliness in work | Disorderly | Somewhat below average in orderliness | Average orderliness | Somewhat above average orderliness | Exceptionally orderly |

In another type, there are lists of statements to be checked but no line with its five or more divisions:¹

The scale describes a set of situations related for the most part to the social adjustment of the pupil. Samples of the situation are as follows:

- I. Involves taking turns on apparatus or in group discussion.
- IV. Child has a social task to be completed.
- VII. Child faced with failure.
- XIII. When things must be organized for work.

Subitems under this last division with weights attached give a more exact idea of how the checking is done.

When things must be organized for work:

Value

- 10 *a.* Gets things he needs together ahead of time so that work goes smoothly.
- 6 *b.* Careful but slow in getting things together.
- 4 *c.* Careless in getting things together.
- 3 *d.* Only gets things as needed.
- 1 *e.* Waits for others to get things for him.

The author has described these scales as follows.²

These samples of rating scales exemplify the leading characteristics of these instruments. All three rating scales are carefully pre-

¹ Van Alstyne, Dorothy, "A New Scale for Rating School Behavior and Attitudes in the Elementary School," *Journal of Educational Psychology* (1936) 27:677-693. Quoted in Jordan, A. M., *Educational Psychology*, 3d ed., p. 565. New York: Henry Holt and Company, Inc., 1942. By permission of Henry Holt and Company, Inc.

² Jordan, A. M., *Educational Psychology*, 3d ed., pp. 562-563. New York: Henry Holt and Company, Inc., 1942. By permission of Henry Holt and Company, Inc.

pared. The traits to be rated are accurately, even meticulously, defined. The division points are made clearly apprehensible by means of words signifying different amounts of the traits being rated. In the best of these scales demarcation points are not blurred by means of some ubiquitous selection of words such as *little*, *fair*, *average*, as applied to the traits being rated but rather are made to stand out by words indicating a certain nicety of distinction such as *defiant*, *critical of authority*, *ordinarily obedient*, etc. These exactly descriptive expressions aid the rater in recognizing the differences otherwise impossible to distinguish.

There are other characteristics of good rating scales apparent in these samples. Scores placed along these lines may be given quantitative aspects simply by designating the first division as "1"; the second as "2"; etc. These numerical records give opportunity for combining one individual's scores on several traits. In the third place, there is a tendency under the influence of careful selection of traits and their accurate description to make the judgments themselves more analytical so that gross total characteristics are broken up into much smaller traits. Finally, you will notice that the material of the scales is given a permanent form in printing, thus again emphasizing the care utilized in their construction.

These excerpts from rating scales illustrate the variation upon a central theme which can be made. In general, they follow the rules of good scale construction pretty closely:

1. Not more than seven divisions of the line
2. Divisions reinforced by careful verbal descriptions
3. A continuous line
4. Simplicity of administration
5. Extremes not so far distant from the mean that nobody will use them
6. Descriptive terms easily understood by the rater

They fail in one recommendation which is worth considering. There is a tendency for ratings to be made near the average when there is doubt and uncertainty, consequently the division "about average" gets such a large number of ratings that they are unwieldy for statistical as well as for practical purposes. For this reason it has been recommended that the two divisions between the median and the extremes in a five-point scale be placed nearer the median than to the extremes. This would make the line look something like this:



In the third type the degrees of amounts represented on the scale are carried in the rater's mind. In this case 1 may represent the least amount; 3, the middle amount; and 5, the greatest amount. One can thus rate cooperation, honesty, emotional balance, etc. In this case, the rater usually translates those numbers into descriptive phrases of his own which characterize the trait being rated and then puts down the proper number. While something of accuracy is probably lost in this procedure, it offers a practical way to get many ratings of each individual. After an individual had made many ratings using the more complete scales with their verbally described divisions on a 5-inch line, the amount of error made on a method of this kind is very small. The author has used this procedure in rating cooperation, intelligence, etc., with satisfactory reliability.

SAMPLES OF RATING SCALES

Thus far samples of techniques which are used in rating have been presented. To complete this exhibit two or three rating scales in their entirety will be presented.

The Haggerty-Olson-Wickman Rating Schedules are divided into two parts, Schedule A and Schedule B.

Schedule A consists of 15 behavior problems whose weights, differing among themselves, are based on the seriousness and frequency of the problem in question. Every one of the 15 may be rated as "has never occurred," "has occurred once or twice but no more," "occasional occurrence," or "frequent occurrence." Each one of these is weighted differently as follows:¹

| | Has never occurred | Has occurred once or twice but no more | Occasional occurrence | Frequent occurrence | Score |
|---------------------------------------|--------------------------|--|--------------------------|------------------------|-------|
| Disinterest in school work... | 0 | 4 | 6 | 7 | |
| Lying..... | 0 | 4 | 6 | 7 | |
| Temper outbursts..... | 0 | 8 | 12 | 14 | |
| Imaginative lying..... | 0 | 12 | 18 | 21 | |
| Obscene notes, talk, or pictures..... | 0 | 12 | 18 | 21 | |

These items had been selected after a thoroughgoing preliminary investigation involving the extended judgments of 500 or more teachers

¹ Items by permission of World Book Company, Yonkers, N.Y.

and some 30 or 40 mental hygienists about the seriousness of traits when they occur in children at certain ages. The instructions are "Put a cross (X) in the appropriate column after each item to designate how frequently such behavior has occurred *in your experience* with this child. . . . The numbers are to be disregarded in making your record."

The nature of Schedule B has already been indicated in the three samples taken from it on page 485.

The reliability coefficients are reported only for the 35 rated items of Schedule B. The reliability by the split-halves procedure is .92. When a correlation is made between the ratings of different teachers it turns out to be .60, and between one teacher's rating and the average of the ratings of three or four teachers the coefficient is .70. If reliability were computed as with other measures the same rater would rate a group of subjects the second time. This procedure undoubtedly makes the reliability too high on account of the memory factor. When a rater has once rated James Sewell, for example, he will on the second rating give him nearly the same position as he did at first. On the other hand, the correlation between ratings by two different raters are probably too low. The same conditions are not repeated because of (1) the different experiences the two raters have had with the subject, and (2) the differences in the set or attitude of the two raters. It would seem therefore that rerating gives a too high reliability coefficient and ratings by different raters, a too low one. The coefficient somewhere between the two more nearly approximates the truth. The true coefficient in this case falls perhaps between .60 and .92 or in the neighborhood of .75 or .80.

This same difficulty appears in the reliability of all rating scales.

The validity may be measured by correlating the ratings of one rater with a group of four or five raters. "The validity of the Behavior Rating Scale has been studied by means of ratings, clinical cases, and the subsequent histories of children." "A composite score on Schedules A and B correlated .76 with the frequency with which a group of children were referred by teachers and monitors to the office of an elementary school principal." It was also demonstrated that half the cases referred to child guidance clinics fell into the highest 10 per cent (*i.e.*, those with most problem difficulties) according to the ratings of teachers (manual).

The manual of the Haggerty-Olson-Wickman Behavior Rating Schedules warns against ratings which are not followed by an attempt to study the case further and to alleviate or correct the conditions found. Probably its best feature is the careful description of the 15 problems which compose Schedule A. For example, "*Speech difficulties.* Under this heading include stuttering or stammering, the substitution of one sound for another, and aural inactivity, as indicated by pronouncing letters or sounds incorrectly or by slurring letters or sounds."

Norms are provided in the nature of tables of distribution of total scores for Schedules A and B, and percentile ranks based on from 1,065 to 2,867 cases. Furthermore, similar tables and percentiles are furnished for each of the four divisions: intellectual, physical, social, and emotional.

The *Winnetka Scale for Rating School Behavior and Attitudes*¹ is made up of 13 school situations with seven more or less desirable degrees of participation. Here are two illustrations:

IV. When a child has a social task to be completed.

| | 1st | 2nd | 3rd | 4th | 5th |
|---|-----|-----|-----|-----|-----|
| Carries task to completion even by sacrifice of other interests. (10) | | | | | |
| Carries task through by steady effort even though it does not harmonize with special interests. (9) | | | | | |
| Carries task through only when it does harmonize with special interests. (6) | | | | | |
| Carries task through although application is unsteady. (3) | | | | | |
| Drops task—loses interest quickly. (1) | | | | | |
| Tries to escape task by contrary behavior or by shifting jobs. (0) | | | | | |

VII. When faced with failure.

| | 1st | 2nd | 3rd | 4th | 5th |
|--|-----|-----|-----|-----|-----|
| Sees causes of failure and corrects it. (10) | | | | | |
| Tries to get help to overcome difficulty. (9) | | | | | |
| Recovers quickly and plans new activity. (6) | | | | | |
| Shows disappointment but continues activity. (4) | | | | | |
| Is apparently indifferent to failure. (2) | | | | | |
| Becomes discouraged easily—must succeed in order to continue activity. (1) | | | | | |
| Becomes irritable or angry, or cries. (0) | | | | | |

There are 13 such situations to be rated. The columns to the right are for different ratings. The numbers in parenthesis after each statement refers to the decile scores. "They represent the percentage of the children studied who were rated at or below the given level of behavior." These decile ratings do not seem to change much with each grade. The test makers designed the scale for ratings over a period of 3 years, with two ratings a year. By means of multiple correlation the 13 situations are so classified that five dimensions of personality may be more

¹ World Book Company, Yonkers, N.Y.

accurately obtained: level of cooperation, social consciousness, emotional security, leadership, and responsibility. Three situations are combined into one of these dimensions. For example, under "cooperation" come ratings on: (1) taking turns with apparatus or materials or in a group discussion, (2) carrying out a group project, and (3) when facing a social situation involving sacrifice of own interests or needs to those of group. By averaging the three deciles received on each situation composing the dimension a percentile score may be obtained for it. In this manner a profile may be secured for each year of rating with percentile positions on each of the five dimensions.

The rating scale was carefully constructed with preliminary observations and ratings by means of which corrections in language and scoring were made. The final norms were secured from the ratings of some 1,200 children. The reliability based on the rerating of the same children after 2 to 8 weeks was .87. The categories also were fairly reliable, varying in their coefficients from .12 to .82. The correlation between these ratings and ratings secured through the Haggerty Olson Wickman Behavior Rating Schedules was .71.

The Personality Rating Scale for Preschool Children¹ of the Merrill Palmer School consists of items to be checked in nine different dimensions: ascendance submission, attractiveness of personality, compliance with routine, independence of adult affection and attention, physical attractiveness, respect for property rights, response to authority, sociability with other children, and tendency to face reality. It was developed especially for the nursery school and has its reliability computed only for this age although it has been used somewhat with children of school age. Each of these divisions has a list of items which the rater simply checks. These items are descriptive of simple habits. For example, under "Compliance with Routine" appear such items as "acts silly at lunch table," "refuses many foods," "dawdles over routine activity." Percentile scores are available for different age groups. In the ascendance submission category percentiles are available for: (1) months 24 to 47, (2) months 48 to 143 and (3) months 144 to 203.

In Table 19 there is a list of other rating scales.

SUMMARY

Our quest for satisfactory personality inventories has been only partially realized. The intangibility and complexity of the traits have in part prevented their satisfactory analysis. When the total personality complex has been broken up into measurable traits there was no assur-

¹ Roberts, Catherine Ellis, and Rachel Stutsman Ball, "A Study of Personality in Young Children by Means of a Series of Rating Scales," *Journal of Genetic Psychology* (1938) 52:79-149.

TABLE 19. LIST OF RATING SCALES NOT INCLUDED IN TEXT

| Name | Grade | Contents and validity | Reliability | Publisher |
|--|------------------------------------|--|-------------|--|
| Rating Scale for School Habits (E. L. Cornell, W. W. Coxe, J. S. Orleans) | Upper grades and high school | Attention, neatness, honesty interest, initiative, ambition, persistence, reliability, and stability. All nine scales contained on one page. $r = .55$ to $.75$ with school marks | None given | |
| American Council on Education Personality Rating Scale | 9-13 | Before rating make observations of subjects. Report instances that support rater's judgment. The descriptive scale (B) includes five traits: industry, ability to control others, appearance and manner, emotional control, distribution of time and energy. A is a graphic scale | | American Council on Education |
| BEC Personality Rating Scale (Business Education Council) | 7-16 | Rates eight areas of personality: (1) mental alertness, (2) initiative, (3) dependability, (4) cooperativeness, (5) judgment, (6) personal impression, (7) courtesy, and (8) health. Each one broken down. For example, under <i>dependability</i> are placed: (1) trustworthiness, (2) persistence, (3) punctuality, (4) obedience to rules | | Harvard University Press |
| Vineland Social Maturity Scale | Infancy through adulthood | The 117 items of the scale are arranged in order of average age norms and are numbered in arithmetic succession from 1 to 117. The groupings of items at age follow pretty closely the pattern of the Binet tests. User of scale needs training in its use. May compute social ages. Author claims it is not a rating scale | | Training School, Vineland, N.J. |

ance that the traits did not overlap to such an extent that the measurement of one trait was not in fact partially measuring another. And yet progress seemed possible only in analysis.

Two major methods were discovered which offered some chance of securing improved measurements: (1) self-rating or self-report, and (2) ratings by others. In self-rating an individual was asked to disclose his

own reactions to situations carefully planned to indicate the presence of some personality traits. It was thought that a constellation of such situations would indicate the presence of certain dimensions of personality such as dominance-submissiveness or neuroticism. As a consequence, questionnaires or, more technically speaking, inventories were prepared on which an individual could register the amount of such a dimension. But even here difficulties arose such as those which had to do with the willingness or ability of an individual to disclose his inner life. Samples of these inventories were presented which avoided or at least ameliorated the effects of some of these errors.

The second method, that of rating by others, avoided at least the error of self-favoritism but added some of its own: inadequate observation, a failure to define clearly the trait being rated, and errors due to personal bias. So ubiquitous were these errors that at least three raters were necessary for dependable results. Some improvement in rating was achieved by using a continuous line below each of whose division points the amount of a trait was verbally described.

In this area of the measurement of personality traits it is well to emphasize the great necessity of interpreting the findings as tentative and inconclusive. In no other area is the need so great for gathering all the available data about a subject and then introducing the results of inventory or rating scale into the total picture. Under such conditions the results of the inventories and rating scales are invaluable. They furnish, if properly interpreted, capital aids in the interpretation of the total personality.

QUESTIONS AND EXERCISES

1. How are self-inventories constructed?

2. Explain the fundamental difficulties and sources of error in self-inventories.

3. Describe the leading characteristics of the Bernreuter Personality Inventory.

4. Why is the validity of inventories so difficult to determine?

5. What characteristics of the Bell Adjustment Inventory recommend themselves for practical use?

6. *a.* Secure a Bernreuter or Bell Inventory and take it. Answer the questions as honestly as you can. Score and interpret the results. How do these findings agree with your understanding of the presence of these traits in you?

b. Would such a procedure in 100 cases be one method of studying the inventory's validity?

7. How does Bell's Inventory differ from that of Bernreuter?

8. How do you account for the wide use in schools of the California Test of Personality? Do you think the name "test of personality" is a correct description of this instrument? Why?

9. From the statistical point of view, why is it dangerous to depend too much on scores on the various dimensions of personality obtained from this test? What is meant by the overlapping of categories? How is this overlapping measured?

10. List the inventories constructed for use with younger children. What

added difficulties are present in evaluating personality traits with these subjects?

11. Describe the Roger's Test of Personality Adjustment. How do the questions differ from those already described? How was it validated?

12. Name three other instruments for measuring personality traits. What traits does each propose to measure?

13. How does the rating scale differ from the self-inventory? What are the leading characteristics of a good rating

scale? Name and illustrate three types of rating scales.

14. What are the leading sources of error inherent in the rating procedure?

15. Describe the main characteristics of the Hagerty-Olson-Wickman Rating Schedules. Include a discussion of their reliability and validity.

16. To what uses could rating scales be put in a progressive school?

17. To what uses could the Winnetka scale be put? The rating scales of the Merrill-Palmer School?

BIBLIOGRAPHY

Books

BELL, HUGH M.: *The Theory and Practice of Personal Counseling*. Stanford University, Calif.: Stanford University Press, 1939.

BUROS, OSCAR K.: *The Nineteen Forty Mental Measurements Yearbook*, pp. 1198-1245. Highland Park, N.J.: The Mental Measurements Yearbook, 1941.

———: *The Third Mental Measurements Yearbook*, pp. 23-114. New Brunswick, N.J.: Rutgers University Press, 1949.

CRONBACH, LEE J.: *Essentials of Psychological Testing*, Chap. 14, "Self-Report Techniques; Personality," Chap. 20, "Projective Techniques," New York: Harper & Brothers, 1949.

FLANAGAN, J. C.: *Factor Analysis in the Study of Personality*. Stanford University, Calif.: Stanford University Press, 1935.

GREENE, EDWARD B.: *Measurements of Human Behavior*, Chaps. 17, 18, 19, "Modes of Adjustment." New York: The Odyssey Press, Inc., 1941.

SUPER, DONALD E.: *Appraising Vocational Fitness*, Chap. XIX. New York: Harper & Brothers, 1949.

SYMONDS, P. M.: *Diagnosing Personality and Conduct*, Chap. III, "Rating Methods," Chap. IV, "The Questionnaire," Chap. V, "Adjustment Questionnaires." New York: Appleton-Century-Crofts, Inc., 1931.

Articles

DARLEY, J. G.: "Tested Maladjustment Related to Clinically Diagnosed Maladjustment," *Journal of Applied Psychology* (1937) 21:632-642.

ELLIS, ALBERT: "The Validity of Personality Questionnaires," *Psychological Bulletin* (1946) 43:385-440.

FILER, H. A., and L. J. O'ROURKE: "Progress in Civil Service Tests," *Journal of Personnel Research* (1923) 1:484-520.

FLANAGAN, J. C.: "Technical Aspects of Multi-trait Tests," *Journal of Educational Psychology* (1935) 26:641-651.

GUILFORD, J. P., and R. B. GUILFORD: "Personality Factors S. E. and M. and Their Measurement," *Journal of Psychology* (1936) 2:109-127.

HATHWAY, S. R.: "The Personality Inventory as an Aid in the Diagnosis of Psychopathic Inferiors," *Journal of Consulting Psychology* (1939) 3:112-117.

JARVIE, L. L., and A. A. JOHNS: "Does the Bernreuter Inventory Contribute to Counseling?" *Educational Research Bulletin* (1938) 17:7-9.

LANDIS, CARNEY, et al.: "Empirical Evaluation of Three Personality Adjustment Inventories," *Journal of Educational Psychology* (1935) 26:321-330.

ROBERTS, CATHERINE ELLIS, and RACHEL STUTSMAN BALL: "A Study

of Personality in Young Children by Means of a Series of Rating Scales," *Journal of Genetic Psychology* (1938) 52:79-149.

SPEER, G. S.: "The Use of the Bernreuter Personality Inventory as an Aid in the Prediction of Behavior," *Journal of Juvenile Research* (1936) 20:65-69.

STOGDILL, EMILY, and MINNIE E. THOMAS: "The Bernreuter Personality Inventory as a Measure of Student

Adjustment," *Journal of Social Psychology* (1938) 9:299-315.

SUPER, DONALD E.: "The Bernreuter Personality Inventory: A Review of Research," *Psychological Bulletin* (1942) 39:94-125.

VAN ALSTYNE, DOROTHY: "A New Scale for Rating School Behavior and Attitudes in the Elementary School," *Journal of Educational Psychology* (1936) 27:677-693.

PART FOUR

Statistical Methods

CHAPTER 19

Statistical Methods

Throughout this book there has been continuous reference to statistical concepts and statistical procedures. For this reason the treatment here is in the nature of a summary and elaboration of statistical concepts already familiar. More concretely, statistics has been used (1) in the construction of tests, and (2) in the interpretation of results. In constructing and standardizing tests mention has been made of norms, percentile or standard scores, reliability, and validity. In the interpretation of results, if complete use is made of the data, mention must be made of tables of distribution, the accuracy of the results, and the meaning of scores such as percentile or standard scores. A few other miscellaneous concepts such as the standard error of estimate and the formula for interpreting the influence of range on correlation have appeared. For the student to get the best results he must follow point by point the treatment in the text and work out the problems introduced in the exercises at the end of the chapter, as well as answer all the questions there proposed.

The following statistical concepts are developed:

1. Measures of central tendency
 - a. Median and other percentiles
 - b. The arithmetic average or mean
 - c. Mode
2. Measures of dispersion or scatter
 - a. Standard deviation, T-score, and standard scores
 - b. Probable error (P.E.)
 - c. Semi-interquartile range (Q)
 - d. Average or mean deviation mentioned but not computed
 - e. Advantages of standard scores
3. The coefficient of correlation
 - a. Pearson product-moment
 - b. Spearman rank-difference correlation method
4. Interpretation of coefficients of correlation
5. Uses of correlation coefficients
 - a. Reliability

- b. Validity
 - c. Prognosis
 - d. Test construction
6. Sampling standard error of the mean and of the standard deviation.

ASSEMBLING THE DATA

Scores gathered as the result of testing usually appear in a disarranged state. Our first problem is to arrange them in an orderly manner so that they may be inspected as a whole. Here is a set of test scores gathered from a test of word knowledge administered to a class of college students:

| | | | | | | |
|----|----|----|----|-------|------|----|
| 92 | 88 | 97 | 95 | (100) | (58) | 90 |
| 94 | 72 | 91 | 83 | 88 | 83 | 87 |
| 82 | 78 | 64 | 68 | 97 | 95 | 86 |
| 85 | 89 | 77 | 61 | 74 | 59 | 85 |
| 86 | 71 | 95 | 90 | 92 | 62 | 80 |
| 91 | 90 | 66 | 63 | 85 | 71 | 78 |

In order to construct a table of distribution the highest and lowest numbers must be found. The highest number is 100 and the lowest, 58. The difference between these scores, called the *range*, is 42. If we use intervals or steps of 1 there would be 42 steps, which would become somewhat unmanageable and would defeat our desire to inspect the scores of the whole class together. Useful results may be had by arranging them in about ten intervals. It is thus convenient to divide the range by 10. Here, then, 42 is to be divided by 10, which gives us 4, and 4 could be used as the size of the interval. Perhaps, for a clearer demonstration, intervals of 5 can be used. This gives us 10 intervals.¹

Before we actually begin the construction of our table of distribution we must decide on the meaning of the numbers used. Are our numbers *continuous* or are they *discrete*? They are discrete if there are definite gaps between, as one child, two children, etc.; they are *continuous* if the scores are finely divisible so that as they are increased they approach ever so closely the next score. For example, in the measurement of length, 5 inches may be increased by .1, .3, .5 and so on to 5.90, 5.95 or even to 5.999, until the measurement is any desired closeness to 6 inches. *Educational and psychological measurements are usually continuous.* For our purposes let us assume that each number is in the middle of a dis-

¹ A good rule to follow is to use from 10 to 20 steps or intervals. In general, the smaller the interval the more accurate the work. The author of this text prefers about twelve intervals for ordinary work. In that case, we divide the range by 12 which will give us the size of interval. This will result in 12 or 13 intervals

TABLE 20. DISTRIBUTION OF VOCABULARY SCORES
(Computation of the median and percentiles)

| Scores | Tallies | Frequency (f) |
|------------------|---------|---------------|
| 100 (99.5-104.4) | | 1 |
| 95 (94.5-99.4) | | 5 |
| 90 (89.5-94.4) | | 8 |
| 85 (84.5-89.4) | | 9 |
| 80 (79.5-84.4) | | 4 |
| 75 (74.5-79.4) | | 3 |
| 70 (69.5-74.4) | | 4 |
| 65 (64.5-69.4) | | 2 |
| 60 (59.5-64.4) | | 4 |
| 55 (54.5-59.4) | | 2 |
| | | $N = 42$ |

a. Median = 50th percentile = 85.6.

50 per cent of $N = 21$

$$\text{Median} = 84.5 + \left(\frac{21 - 19}{9} \right) 5$$

Start at bottom, $2 + 4 + 2 + 4 + 3 + 4 = 19$. 9 is the number of cases at the interval in which the median falls. 84.5 is the lowest point in the interval in which the median falls.

b. $Q = (Q_3 - Q_1)/2$. Q_3 is the 75th percentile; Q_1 the 25th percentile.

$$Q_3 \text{ (75th percentile)} = 91.69$$

75 per cent $N = 31.50$. Start at bottom, $2 + 4 + 2 + 4 + 3 + 4 + 9 = 28$. $89.5 + [(31.50 - 28)/8]5$. 89.5 is the lowest point in the interval in which Q_3 falls. 8 is the number of cases in the interval in which Q_3 falls.

$$Q_1 \text{ (25th percentile)} = 72.62$$

25 per cent $N = 10.5$

Start at bottom, $2 + 4 + 2 = 8$. $Q = 69.5 + [(10.5 - 8)/4]5$. 69.5 is the lowest point of interval in which Q_1 falls. 4 is the number of cases in the interval in which Q_1 falls.

$$Q = \frac{Q_3 - Q_1}{2} = \frac{91.69 - 72.62}{2} = \frac{19.07}{2} = 9.53$$

tance. For example, 95 stands for 94.5 to 95.49. It might stand for 95 to 95.9 but we shall use the former. A good illustration of this usage occurs in the custom of life-insurance companies in computing age. To them a person 17 years of age is not thought of, as is ordinarily the case, as having reached 17 at his *last* birthday and as reaching 18 when his

seventeenth year is finished. Life-insurance companies compute age to the *nearest* birthday. Age 17, then, extends from 16 years and 6 months to 17 years and 5 months. In brief, 17 is 16.5 to 17.49. This method more nearly approximates the truth than does the computation of age from the last birthday.

Let us now proceed to construct our table. It must extend high enough to include 100 and low enough to include 58. We now start with 100 and drop by steps of 5 to 55, *i.e.*, our table must include both 100 and 58. By definition the 55 stands for 54.5 up to 59.5, 75 stands for 74.5 up to 79.5, 80 stands for 79.5 up to 84.5, etc. At 80 are included scores from 80 through 84, at 85 are included the scores 85 to 89, etc. We now transfer our 42 scores on page 500 to our Table 20. For each score there is a tally entered in the proper place. For score 92 a tally is entered in the table at 90, for score 94 a tally is also entered at 90, for 82 a tally is entered at 80 in the table, etc., until all are entered. The next step is to count up the tallies and enter their number for each interval in the column labeled "frequency" or *f*.

MEASURES OF CENTRAL TENDENCY

The measures of central tendency are median, mean, and mode. Of these three the median and mean (arithmetic average) are used very frequently while the use of the mode is rare.

MEDIAN AND OTHER PERCENTILES

The median is defined as the mid-point in a table of distribution such as Table 20. (If the numbers are not grouped, the median is sometimes taken as the mid-number.) It is evident that the mid-point is also the fiftieth percentile. The procedure for computation is straightforward. We take $\frac{1}{2}$, or 50 per cent, of the number of cases (*N*) and then discover how far up the scale this number extends. By observing Table 20 we find *N* is 42. $N/2$ is therefore $42/2$ or 21 (the half sum). We now begin at the *bottom* of the frequency column and count *up* until we come to the interval in which case 21 falls, as follows: $2 + 4 + 2 + 4 + 3 + 4$, the sum of which is 19. We still need two more cases to arrive at 21, the half sum. These two cases are taken out of the nine cases at the next interval. It is assumed that the nine cases are evenly distributed over step 85. The median then is 84.5 (the beginning of this interval of 85) $+ (2/9)5 = 85.61$. We multiply $2/9$ by 5 because 5 is the size of the interval. In brief, this process becomes $84.5 + [(21 - 19)/9]5 = 85.61$. It is seen that in this way the mid-point, the median, is discovered. The median is frequently used because it is easy to compute and is not greatly affected by extreme scores.

Percentiles

It was pointed out in computing the median that it was the 50th percentile. In computing the median we simply compute $\frac{1}{2}$ or 50 per cent of the scores and discover where this number falls in the table of distribution. Exactly the same procedure is used in computing any percentile. We take the percentage of the cases we desire and discover by interpolation its exact location. Thus for the 10th percentile we take 10 per cent of the cases and interpolate, for the 20th percentile we take 20 per cent and interpolate, etc. In this manner it is possible to compute any percentile from 1 to 100.

Computation of Percentiles

To compute the 15th percentile, take 15 per cent of N , here 42 (Table 20). This is 6.3. Count up from the bottom of the frequency column until you come to the interval in which 6.3 ends. In Table 20 this becomes $2 + 4$ and .3 is left over. The 15th percentile is then $64.5 + [(6.3 - 6)/2]5 = 65.25$. The 6 in the numerator is the sum of the cases below interval 65. The number 64.5 is the lowest point in the interval 65. The number 2 indicates the number of cases at interval 65.

To compute the 65th percentile, take 65 per cent of 42, which is 27.3. Count up the frequencies in Table 20 until the next step contains the last of 27.3 as follows: $2 + 4 + 2 + 4 + 3 + 4 = 19$. Now there are 8.3 cases left which are contained in the 9 at 85. Computing, $84.5 + [(27.3 - 19)/9]5 = 84.5 + 4.61 = 89.11$. Thus we see 84.5 is the lowest point of the interval in which the 65th percentile falls, 27.3 is 65 per cent of 42, 19 is the sum of the cases up to interval 85, and 9 are the cases evenly distributed over 85. You may check your understanding of the procedures by comparing your computations with the following answers: 40th percentile = 81.75; 70th percentile = 90.37; 1st percentile = 55.55; 25th percentile = 72.62; and 75th percentile = 91.69.

Percentiles furnish points of reference in the norms of a large number of tests. When tests were first standardized the usual percentiles computed were the 25th, 50th, and 75th. But, as experience increased, the need was felt for further points of comparison, *i.e.*, percentile points, all up and down the line. In interpreting such percentile points one must remember that the 25th percentile simply means that 25 per cent of the cases are below that point while 75 per cent are above it, and that the eighty-fifth percentile means that 85 per cent are below that point and 15 per cent above it.

THE ARITHMETIC AVERAGE OR MEAN

The most familiar measure of central tendency is the arithmetic average or the mean. It is computed by adding up the quantities and

dividing the sum by their number. Here we could add up our numbers and divide the sum by 42. When the data are grouped as in our Table 21, the mean may be computed by first assuming the mid-point of some interval as the mean and then adding or subtracting the proper correction, thus arriving at the mean. Table 21 indicates the process.

TABLE 21. COMPUTATION OF THE MEAN

| Mid-points | | <i>f</i> | <i>d</i> | <i>fd</i> |
|------------|------------------|---------------|----------|-----------|
| 102 | 100 (99.5-104.4) | 1 | 4 | 4 |
| 97 | 95 (94.5-99.4) | 5 | 3 | 15 |
| 92 | 90 (89.5-94.4) | 8 | 2 | 16 |
| 87 | 85 (84.5-89.4) | 9 | 1 | 9 |
| 82 | 80 (79.5-84.4) | 4 | | |
| 77 | 75 (74.5-79.4) | 3 | -1 | -3 |
| 72 | 70 (69.5-74.4) | 4 | -2 | -8 |
| 67 | 65 (64.5-69.4) | 2 | -3 | -6 |
| 62 | 60 (59.5-64.4) | 4 | -4 | -16 |
| 57 | 55 (54.5-59.4) | 2 | -5 | -10 |
| | | <i>N</i> = 42 | | |

Mean = assumed mean + Ci (correction \times interval)

$$C \text{ (correction)} = \frac{\sum fd}{N} = \frac{44 - 43}{42} = \frac{1}{42} = .024$$

i (interval) = 5

$$\text{Mean} = 82 + (.024)5 = 82.12$$

In the computation of the mean from grouped data we may proceed by (1) the long method, or (2) the short method. The answers are exactly the same in both cases. In the long method the mid-point of each interval is multiplied by the frequency. The sum of these products is secured and divided by N . This will give us the mean from grouped data. In this instance: 102×1 , 97×5 , 92×8 , etc. The sum of the products for the 10 intervals is 3,449, which when divided by 42 gives us 82.12. In the *short method* a mean is assumed, deviations are computed from this assumed mean and multiplied by the proper frequency at each interval. The algebraic sum of these products is taken, divided by the number of cases, then multiplied by the size of the interval and added algebraically to the assumed mean. Table 21 shows that the assumed mean is 82. Column d contains simply the number of steps above (+) or below (−) the assumed mean. Column fd is, of course, the product of column f and column d . The computation and answer, 82.12, appear in Table 21.

In this problem you will note that the median is 85.6 while the mean

is 82.1. The mean was pulled down by the six cases at 55 and 60. Extreme scores are weighted according to their size in computing the median. *It is well to remember that extreme cases affect the mean more than they do the median.*

THE MODE

The *mode* may be thought of as the "value in a series at which the greatest frequency lies." This value, as may be seen in Table 21, is 87. The mode is also calculated from the formula:

$$\text{mode} = (3 \times \text{median}) - (2 \times \text{mean}).$$

In our problem this would be $(3 \times 85.6) - (2 \times 82.1)$ or 92.6. This does not seem to be a very representative number for our distribution.

MEASURES OF DISPERSION OR SCATTER

Two questions stand preeminent when a table of distribution is in question: (1) what is its central tendency? (2) What is its dispersion or scatter? The second question may be asked more informally: How closely around the central tendency are the cases grouped? Are they packed in close, or are they scattered out until there is no semblance of unity in the group studied? There are four of these measures which differ in quantity but not in quality, *i.e.*, which differ as the meter differs from the yard:

1. Standard deviation (S.D.)
2. Probable error (P.E.)
3. Semi-interquartile range (Q)
4. Average or mean deviation (A.D.)

Under normal conditions the P.E. and Q are equal. The standard deviation is larger than the others (P.E. = 0.6745 S.D.).

STANDARD DEVIATION

In a normal or symmetrical curve the standard deviation is the distance from the mean which in one direction includes about 34 per cent (34.13) of the total cases. Sometimes this distance out from the mean is large, in which case the members of the population are unlike each other; sometimes it is small, which indicates closer resemblances among members of that group or population. A class with a large standard deviation in intelligence would be heterogeneous; one, with a small standard deviation, homogeneous *i.e.*, with respect to intelligence. Table 22 shows how the standard deviation is computed.

In computing a standard deviation the same procedure is used as in computing the mean. Additional work is needed to compute the fd^2

TABLE 22. COMPUTATION OF THE STANDARD DEVIATION

| Scores on word knowledge | <i>f</i> | <i>d</i> | <i>fd</i> | <i>fd</i> ² |
|--------------------------|---------------|----------|-----------|------------------------|
| 100 | 1 | 4 | 4 | 16 |
| 95 | 5 | 3 | 15 | 45 |
| 90 | 8 | 2 | 16 | 32 |
| 85 | 9 | 1 | 9 | 9 |
| 80 | 4 | | | |
| 75 | 3 | -1 | -3 | 3 |
| 70 | 4 | -2 | -8 | 16 |
| 65 | 2 | -3 | -6 | 18 |
| 60 | 4 | -4 | -16 | 64 |
| 55 | 2 | -5 | -10 | 50 |
| | <i>N</i> = 42 | | | 253 |

$$\text{Sum of } fd^2 = \Sigma fd^2 = 253$$

$$C = \frac{\Sigma fd}{N} = \frac{1}{42} = .02 \text{ (or } .024)$$

$$i = \text{interval} = 5$$

$$\begin{aligned} \text{S.D.} &= \left[\sqrt{\frac{\Sigma fd^2}{N} - C^2} \right] i \\ &= [\sqrt{253/42 - (.02)^2}] 5 \\ &= 2.45 \times 5 = 12.25 \end{aligned}$$

column and then to substitute in the formula. If our curve were normal, the mean, plus or minus 1 standard deviation, would include 68.26 per cent of the total. In our case the mean is 82.1 ± 12.25 . If we add 12.25 to 82.1 we get 94.35, and if we subtract 12.25 from the mean the score is 69.85. Between these limits there are 28 cases or about 67 per cent of the total.

THE PROBABLE ERROR (P.E.)

In a normal curve a probable error equals 0.6745 times the standard deviation. The probable error is so frequently used in this manner that we shall not introduce other ways of computing it.

THE SEMI-INTERQUARTILE RANGE (Q)

The formula for the interquartile range is $Q = (Q_3 - Q_1)/2$ in which Q_3 , the third quartile, is our old friend the 75th percentile and Q_1 , the first quartile, is the 25th percentile. A little thought shows that $Q_3 - Q_1$ gives us the middle 50 per cent of the cases. If we divide that by two, we have the 25 per cent of cases on either side of the median. It is so

easily computed that it has been frequently used and sometimes in place of the probable error (see Table 20).

THE AVERAGE OR MEAN DEVIATION

This measure is computed by averaging the deviations from the mean regardless of signs. If 25 students guessed the time of day they would miss the true time by varying amounts in either direction. If we simply average these deviations we would have the average deviation although the point of reference is the *mean* rather than a true score such as the one we have used here.

USES OF STANDARD DEVIATION

One use of the standard deviation is of the greatest importance. It is the so-called *standard score*. The formula is: standard score = $(X - M_x)/\sigma_x$,

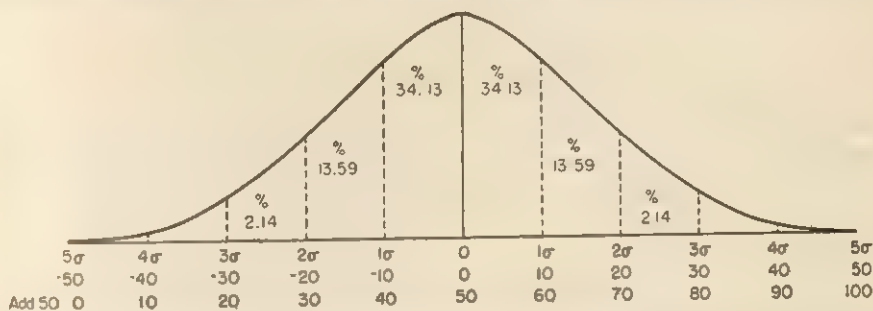


FIG. 38. Normal curve, sigma units. Percentage in each sigma value. Bottom line, T-scores.

where X is a single person's score, M_x is the mean, and σ_x the standard deviation (sigma). One member of our group scored 78. Substituting in the formula we get:

$$\text{standard score} = (78 - 82.1)/12.25 = -4.1/12.25 = -.3.$$

The score of 78, then, is only 0.3 standard deviation units *below* the mean of the group.

Arising out of the concept of the standard score is the T-score. Originally this idea came from an attempt to develop units of mental measurement which would be equivalent. Standard-deviation units derived from a representative group of 12-year-olds (McCall) were treated as shown in Fig. 38. Each sigma distance was divided into 10 equal parts. This gave a set of scores ranging from -50 through 0 to +50. Negative scores are always troublesome. To get rid of them McCall assumed a mean of 50 which, when added all along the line, gave a series of numbers beginning with 0 and going to 100. These numbers were convenient and

most important of all they were about equal to each other. A change from 20 to 30 is nearly equal to a change from 40 to 50, or from 70 to 80. In brief, these units are about the best we have. This procedure has been generalized into a formula. $T\text{-score} = 50 + [(X - M)/\sigma] 10$ in which X is an individual's score, M is the mean, and σ the usual standard deviation. This formula is accurate only when the distribution is normal but works fairly well even when the original distribution deviates slightly from the normal. Let us take our distribution based on 42 cases (Table 22) which has a mean of 82.1 and a standard deviation of 12.25. Applying the formula we get $T\text{-score} = 50 + [(X - 82.1)/12.25]10$. What would be the T-score of a person who scored 92?

$$T\text{-score} = 50 + [(92 - 82.1)/12.25]10 = 50 + [(9.9/12.25)]10 = 58$$

If we take another actual score 58, the lowest case, we get

$$T\text{-score} = 50 + [(58 - 82.1)/12.25]10 = 31$$

Some test constructors have preferred to use a standard deviation of 20 and a mean of 100. This gives a range from 0 to 200. You will see in most tests recently constructed a little table at the end of each test by which raw scores (or simply test scores) can be transmuted into standard scores, or T-scores.

ADVANTAGES OF STANDARD SCORES

The two most popular procedures for changing raw scores which differ largely in meaning to equivalent scores are (1) the standard score, and (2) the percentile. The percentile has the advantage of being easily understood. A percentile score of 60 means that this is the position in 100—that 60 per cent of the cases are below the score in question and 40 per cent are above. If a percentile score is 32, then 32 per cent are below it and 68 per cent are above it. Standard scores have no such clarity of understanding. A standard score (T-score) of 60 based on a mean of 50 and an S.D. of 10 would mean that this case is 1 standard deviation above the mean. If this score were to be transmuted into percentiles it would be in the 84th percentile ($50 + 34$, since in a normal curve 1 standard deviation includes 34 per cent on either side of the mean). From the standpoint of accuracy the standard score has great advantages.

Consider Fig. 39, which illustrates the differences between these two measures. The standard score is based on the assumption of normal distribution. While the units along the base line are equal, the percentage of cases included under each unit increases greatly as we approach the mean and decreases as we go past the mean to the extreme scores. For example, one sigma nearest the mean includes 34.13 per cent in a normal

distribution, while only a little more than 2 per cent of the cases appear between the second and third standard deviations. In short this measure allows for that queer arrangement of actual scores known as the normal distribution. The real distance between the highest 1 per cent in intelligence, for example, and the next is far more than that between the 49th and 50th percentiles. The percentile assumes a rectangular distribution, as in *b* (Fig. 39). The percentages above each percentile are assumed to be the same all along the base line which *simply is not a fact* in the usual collection of data.

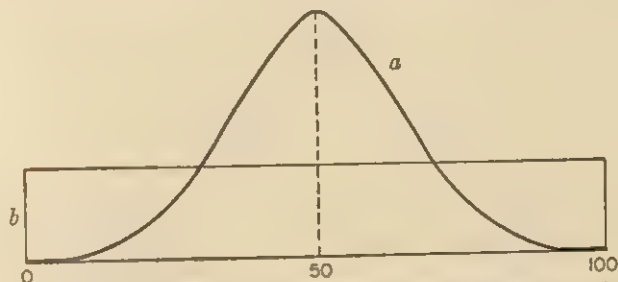


FIG. 39. Curve showing difference between percentiles and standard scores.

The percentile works very well between the 25th and the 75th percentile but errs greatly in the extremes.

THE COEFFICIENT OF CORRELATION

Thus far we have been speaking of the statistics involving one variable. The 42 scores studied were secured from a vocabulary test. Each individual had just *one score*. In correlation, on the other hand, in most of our work there are always two measures for each subject. The problem is to discover the mutual relation, the correlation, between these measures. We have been discussing correlation since our study of reliability and validity. The index of reliability is usually a correlation between two forms of a test, the repetition of the same test, or the odd scores against the even scores. It was there indicated that reliabilities above .90 were highly desirable. Correlation might be defined as the *average degree of resemblance which exists between two tests, or two traits in the same group of individuals, each individual being measured twice*. It must be realized that other facts may be correlated which have no direct relation to human measures such as the correlation between average rainfall and crop yield. For our purposes, correlation will usually be computed between two measures of human traits and there will be a considerable number of individuals measured or else the coefficient will not be reliable.

TABLE 23. COMPUTATION OF THE COEFFICIENT OF CORRELATION*

| Word knowledge, X | Miller, Y | x | y | x ² | y ² | xy |
|----------------------|--------------|-------|-----|-------------------------|-------------------------|-------------|
| 92 | 89 | 9 | 3 | 81 | 9 | 27 |
| 88 | 86 | 5 | 0 | 25 | 0 | 0 |
| 97 | 114 | 14 | 28 | 196 | 784 | 392 |
| 95 | 104 | 12 | 18 | 144 | 324 | 216 |
| 100 | 117 | 17 | 31 | 289 | 961 | 527 |
| 58 | 58 | -25 | -28 | 625 | 784 | 700 |
| 90 | 114 | 7 | 28 | 49 | 784 | 196 |
| 94 | 105 | 11 | 19 | 121 | 361 | 209 |
| 72 | 82 | -11 | -4 | 121 | 16 | 44 |
| 91 | 76 | 8 | -10 | 64 | 100 | -80 |
| 83 | 102 | 0 | 16 | 0 | 256 | 0 |
| 88 | 92 | 5 | 6 | 25 | 36 | 30 |
| 83 | 65 | 0 | -21 | 0 | 441 | 0 |
| 87 | 78 | 4 | -8 | 16 | 64 | -32 |
| 82 | 103 | -1 | 17 | 1 | 289 | -17 |
| 78 | 62 | -5 | -24 | 25 | 576 | 120 |
| 64 | 76 | -19 | -10 | 361 | 100 | 190 |
| 68 | 62 | -15 | -24 | 225 | 576 | 360 |
| 97 | 109 | 14 | 23 | 196 | 529 | 322 |
| 95 | 95 | 12 | 9 | 144 | 81 | 108 |
| 86 | 69 | 3 | -17 | 9 | 289 | -51 |
| 85 | 78 | 2 | -8 | 4 | 64 | -16 |
| 85 | 96 | 2 | 10 | 4 | 100 | 20 |
| 89 | 104 | 6 | 18 | 36 | 324 | 108 |
| 77 | 78 | -6 | -8 | 36 | 64 | 48 |
| 61 | 64 | -22 | -22 | 484 | 484 | 484 |
| 74 | 82 | -9 | -4 | 81 | 16 | 36 |
| 59 | 58 | -24 | -28 | 576 | 784 | 672 |
| Sum (Σ) | 2,318 | 2,418 | | Σx ² = 3,938 | Σy ² = 9,196 | Σxy = 4,613 |
| Mean (M) | 83 | 86 | | | | |

$$\begin{aligned}
 r &= \frac{\Sigma xy}{\sqrt{\Sigma x^2} \sqrt{\Sigma y^2}} \\
 &= \frac{4,613}{\sqrt{3,938} \sqrt{9,196}} \\
 &= .77
 \end{aligned}$$

* From Jordan, A. M., *Educational Psychology*, 3d ed., p. 473. New York: Henry Holt and Company, Inc., 1942. By permission.

PEARSON PRODUCT-MOMENT METHOD

To Sir Francis Galton is usually given the honor of having first developed and used the coefficient of correlation as we know it. It was Karl Pearson, of the University of London, who derived for us the mathematical formula. The formula is $r = \Sigma xy / N \sigma_x \sigma_y$ in which r is the coefficient of correlation, x and y are deviations from their respective means and are the same as d as we have used it, Σ is the sum (after the deviations have been multiplied), N is the number of pairs, σ_x is the standard deviation (S.D.) of one variable, and σ_y is the standard deviation (S.D.) of the other. In the definition the term "average" was used. This term can be better understood if you note the N in the denominator. It is well to relate this formula a little more closely to what has already been learned in statistics. The standard score is $(X - M_x) / \sigma_x$ in which X is a score and M_x is a mean. Now $X - M_x = x$, or d as we have used it. We now get, by substitution, x for $X - M$, in the formula for the standard score, x / σ_x . In like manner the standard score for y is y / σ_y . Now by multiplying these standard scores together and adding up the xy products and multiplying the products of the standard deviations by the number of pairs, we obtain the coefficient of correlation.

Examine the following computation very carefully both to learn how to perform it and to understand it.

You will note that the capitals X and Y represent individual scores (Table 23). The small letters x and y represent deviation from the means, here taken as the nearest whole numbers. For example, John received 92 on word knowledge (X) and 89 on Miller Intelligence Test (Y). The mean of the word knowledge scores (M) is 83; that for Miller (MY), 86. Small x then is $92 - 83 = 9$; small y , $89 - 86 = 3$. In like manner for the second pair, $X - Mx = x$, or $88 - 83 = 5$, and also $Y - My = y$, or $86 - 86 = 0$, etc. From now on it is simply a process of substituting in the formula $r = \Sigma xy / N \sigma_x \sigma_y$. For the σ_x we substitute its equal, $\sigma_x = \sqrt{\Sigma x^2 / N}$ and for $\sigma_y = \sqrt{\Sigma y^2 / N}$. For the total formula we now have

$$r = \frac{\Sigma xy}{N \sqrt{\Sigma x^2 / N} \sqrt{\Sigma y^2 / N}}$$

Now, $\Sigma xy = 4,613$, $\Sigma x^2 = 3,938$, $\Sigma y^2 = 9,196$, and $N = 28$. Therefore we have

$$r = \frac{4,613}{28 \sqrt{3,938/28} \sqrt{9,196/28}}$$

The 28's cancel out, for $28 \sqrt{1/28 \times 1/28} = 28/28 = 1$. Therefore

$$r = \frac{4613}{\sqrt{3,938} \sqrt{9,196}} = .766 \text{ (or .77)}$$

Please note that this is the coefficient when we use for the mean its nearest whole number. The mean of the word knowledge scores is 2,318 divided by 28 or 82.78 and the mean for the Miller scores is 2,418 divided by 28, or 86.36. There are ways of correcting for this use of the nearest whole number for the mean, but in most cases the difference in r is negligible. In this case the r when computed from the means of 82.78 and 86.36 is .767.

SPEARMAN'S RANK-DIFFERENCE CORRELATION METHOD

The method of rank differences is a somewhat simpler way of computing the coefficient of correlation. It is only a trifle less exact and can be converted into the Pearsonian r by means of tables. The coefficient is called *rho* (ρ) to distinguish it from the Pearson r . The differences between the two coefficients is hardly ever more than .02 and is often nearer .01. There are several occasions when it makes a definite contribution:

1. When the scores themselves are gathered in the forms of ranks such as the ranking of a class for honesty or for cooperativeness. For example, the problem of whether cheating is correlated with cooperativeness.

2. When the number of scores is small and a quick answer is wanted. This method is rarely used when the number of pairs in the computation is more than 50.

In our illustration of this procedure (Table 24) the same scores are used which were utilized in the computation of r . In general, the procedure is to rank the scores in each variable (here X and Y) subtract the ranks and place the differences under the column marked d , and then square these differences (d^2). The rest is simply a matter of adding up the d^2 and substituting in the formula $\rho = 1 - \frac{6\sum d^2}{N(N^2 - 1)}$ where d is the difference in ranks and N is the number of pairs.

Let us look now at the process exemplified in Table 25. You will note that there are 28 pairs as before (Table 24). We have then a part of our correlation already. The denominator of our fraction becomes $28(784 - 1)$ when 28 is substituted for N in the formula $N(N^2 - 1)$. What now remains to be done is to compute the numerator of the fraction.

In computing the numerator we first rank the numbers in each column. In column X the largest number is 100, so its rank is 1. The

TABLE 24. COMPUTATION OF ρ BY THE METHOD OF SQUARED DIFFERENCES IN RANK

| Word knowledge, X | Miller, Y | Rank X | Rank Y | d^* | d^2 |
|----------------------|--------------|-----------|-----------|-------|--------|
| 92 | 89 | 7 | 13 | 6 | 36 |
| 88 | 86 | 11.5 | 14 | 2.5 | 6.25 |
| 97 | 114 | 2.5 | 2.5 | | |
| 95 | 104 | 4.5 | 6.5 | 2 | 4 |
| 100 | 117 | 1 | 1 | | |
| 58 | 58 | 28 | 27.5 | .5 | .25 |
| 90 | 114 | 9 | 2.5 | 6.5 | 42.25 |
| 94 | 105 | 6 | 5 | 1 | 1 |
| 72 | 82 | 23 | 15.5 | 7.5 | 56.55 |
| 91 | 76 | 8 | 20.5 | 12.5 | 156.25 |
| 83 | 102 | 17.5 | 9 | 8.5 | 72.25 |
| 88 | 92 | 11.5 | 12 | .5 | .25 |
| 83 | 65 | 17.5 | 23 | 5.5 | 30.25 |
| 87 | 78 | 13 | 18 | 5 | 25 |
| 82 | 103 | 19 | 8 | 11 | 121 |
| 78 | 62 | 20 | 25.5 | 5.5 | 30.25 |
| 64 | 76 | 25 | 20.5 | 4.5 | 20.25 |
| 68 | 62 | 24 | 25.5 | 1.5 | 2.25 |
| 97 | 109 | 2.5 | 4 | 1.5 | 2.25 |
| 95 | 95 | 4.5 | 11 | 6.5 | 42.25 |
| 86 | 69 | 14 | 22 | 8 | 64 |
| 85 | 78 | 15.5 | 18 | 2.5 | 6.25 |
| 85 | 96 | 15.5 | 10 | 5.5 | 30.25 |
| 89 | 104 | 10 | 6.5 | 3.5 | 12.25 |
| 77 | 78 | 21 | 18 | 3 | 9 |
| 61 | 64 | 26 | 24 | 2 | 4 |
| 74 | 82 | 22 | 15.5 | 6.5 | 42.25 |
| 59 | 58 | 27 | 27.5 | .5 | .25 |

$$N = 28$$

$$\Sigma d^2 = \text{sum of } d^2 = 816.50$$

$$\rho = 1 - \frac{6\Sigma d^2}{N(N^2 - 1)}$$

$$= 1 - \frac{6(816.50)}{28(783)} = 1 - \frac{4,899}{21,924}$$

$$= 1 - .223$$

$$= .777(\text{or } .78)$$

* d = difference in ranks. Since d is squared, signs are always plus.

number next in size is 97, but you will note there are two of them. Each has an equal right to be ranked 2 and the other would then, of course, be ranked 3. What we do is simply take the mean of the ranks and give each one 2.5. Notice that we have now used ranks 1, 2, and 3 and that the next number will be ranked 4. The number next in size is 95, but there are two of them; hence we give each one 4.5. We have now used up ranks through 5 so that 94, the number next in size, is ranked 6. A good check is to make sure that your lowest rank is the same as the number of pairs. If you count wrong your last rank will not equal N , the number of pairs. In our case 58 is the smallest number in column X , and its rank is 28. We proceed in exactly the same way in column Y . Looking down this column, we find 117 the highest score, so we label it 1. There are two scores of 114, hence each is given 2.5. You will note in column Y that there are three scores of 78. Since 16 ranks had been used up when this score appeared, the 17, 18, and 19 ranks would be used with these three numbers. The mean of these three is 18, consequently each 78 is given a rank of 18. Once the rankings are made, simply subtract them pair by pair without regard to signs, square the d 's, add them up, and substitute in the formula.

INTERPRETATION OF COEFFICIENTS OF CORRELATION

A coefficient of correlation is dependent for its meaning on (1) its size, and (2) the size of the sample and its representativeness of the population from which it was drawn. For example, 18-year-old college students would not be representative of the total population of 18-year-olds.

Size of the Coefficient

In general, the nearer the coefficient is to $+1.00$ or -1.00 , the higher the correlation and the closer the resemblance. We have said that reliability coefficients should usually be .85 or above in most cases. When the coefficient approaches zero, let us say when it varies from $-.15$ to $+.15$, it is no larger than would occur by chance, *i.e.*, there is no relationship between the two variables studied. When the coefficient is .20 and above, its meaning and value depends upon its relation to other variables. If, let us say, a coefficient of .23 is computed with a criterion and if this variable is not related to the other factors in a test battery it may be profitable to use it. Correlations in the neighborhood of .50 or .60 have been called "marked," "significant," and under some conditions, "high." A *correlation* of .60 between intelligence-test scores and students' marks would be *high*. On the other hand, a *reliability* of .75 would be definitely *low*. You see, the interpretation of the coefficient is partly a matter of magnitude and partly a matter of the type of relation which it expresses.

Reliability of the Coefficient

One of the problems which always confronts an investigator is whether this correlation which he has computed is representative or not. In technical terms is the computed r a true r ? It must always be kept in mind that the pairs drawn are only a sample of what the total population is. The data with which the two methods of correlation were illustrated were (X) scores on an intelligence test and (Y) scores on a test of word knowledge. These were drawn from a college population. The true correlation would be that computed from the use of scores secured from all college students. Fortunately the correlation between the 28 pairs drawn at random give some indication as to what the true r would be. The formula for the standard error of the Pearson coefficient of correlation is $S.E._r = (1 - r^2)/\sqrt{N - 1}$. Clearly, its size depends on the size of r and the size of N . If N is very large, the fraction is small and $S.E._r$ is small, a condition which indicates high reliability. If r is large and N is large, the r is very reliable. If we use our coefficient we get $S.E._r = (1 - r^2)/\sqrt{N - 1} = (1 - .593)/5.196 = .078$ (or .08). We may now write $r = .77 \pm .08$, which when interpreted means:

1. The chances are 68 in 100 (see page 507) that the true r lies between .69 and .85. This is 1 standard error limit.
2. The chances are 95 in 100 that the true r lies between plus or minus 2 $S.E._r$, or between .61 and .93.
3. The chances are 99.7 in 100 that the true r lies between plus or minus 3 $S.E._r$, or between .53 and 1.00.

The numbers 68, 95, and 99.7 are taken from a table which shows the percentage of total scores appearing under a graph representing the normal curve at 1 S.D. (or here $S.E.$), 2 S.D.s, and 3 S.D.s.¹ It is thus seen that while the true r cannot be calculated, its limits can.

USES OF THE COEFFICIENT OF CORRELATION

The coefficient of correlation is one of the most widely used statistical concepts. Not only in the fields of education and psychology has it achieved great statistical prominence but also in the fields of agriculture, sociology, and economics, to name a few, it has found favor. In testing, this concept has been useful in four areas: (1) reliability, (2) validity, (3) prognosis, and (4) test construction.

Reliability

In computing the reliability of tests the coefficient of correlation is almost universally used. Whether the reliability is computed by the

¹ See Garrett, H. E., *Statistics in Psychology and Education*, 3d ed., p. 115. New York: Longmans, Green & Co., Inc., 1947.

repetition of the same test, by the administration of two forms of the same test, or by the odd-even technique, correlation is used. The symbols are usually r_{11} for repetition, r_{AB} for two forms and r_{11} for the odd-²¹¹

even technique (see page 28). The reliabilities of batteries of achievement tests usually run .95 or above those of intelligence tests only slightly lower, and those of inventories, about .85 to .92. The higher the coefficient, the less variation is there from one form to the other, *i.e.*, the more accurate is the test. The reliabilities of school marks, except where the testers are trained, are in the neighborhood of .65 to .75. An extension of the notion of reliability appears in the standard error of a score or, as it is usually named, the *standard error of measurement*. When an individual receives a score on a test this number is *not the true score*. It is a sample score. We might assume that if such a one were tested a thousand times on such a test one could obtain a true score. What we have, then, is a sample from which we can predict a true score. The formula for a standard error of measurement is:

$$S.E._{meas.} = \frac{\sigma_1 + \sigma_2}{2} \sqrt{1 - r}$$

If the two standard deviations are equal the formula becomes

$$S.E._{meas.} = \sigma \sqrt{1 - r}.$$

From the formula it is clear that the amount of variation expected from a single score depends on (1) the standard deviations of the two variables being studied, and (2) the *size* of the coefficient of correlation. If r were 1.00, the variation would be 0. By the use of this formula it is possible to predict within what limits the true score most probably lies. In computing the reliability of a test suppose that the correlation between two forms were .96 with a standard deviation of 10; then $S.E._{meas.} = 10 \sqrt{1 - .96} = 2$. If on this test one subject receives a score of 65, then we could say that the chances are 68 in 100 that the true score lies between 63 and 67 (1 S.D.), 95 in 100 that the true score lies between 61 and 69 (2 S.D.s) and more than 99 in 100 that the true score lies between 59 and 71 (3 S.D.s). These numbers are secured from a table which indicates the percentage under different portions of a normal curve. It has been demonstrated that deviations from a true score fall into the form of the normal curve.

To return to our $S.E._{meas.}$, we find that it is easily understood and an excellent measure of reliability. Suppose that instead of a S.D. of 10 and a correlation of .96 between the two test forms the S.D. had been 15 and the reliability coefficient .85. Substituting in our formula, we get

$S.E._{meas.} = 15 \sqrt{1 - .85} = 15 \times .39 = 5.84$. Let us round off this 5.8 and call it 6. Our reliability now becomes 65 ± 6 . The chances are 68 in 100 that the true score lies between 59 and 71 (1 S.D.); 95 in 100 that it lies between 53 and 77 (2 S.D.s); and more than 99 in 100 that it lies between 47 and 83. It is easily seen that, if we have to go as low as 47 and as high as 83 to get the true score, our sample score is not of much value. In the first instance, with an S.D. of 10 and a correlation of .96, the true score had an extreme variation of 59 to 71. It is clearly seen that one needs a high reliability in a test if it is to be of any real use for individual diagnosis.

Another use of the reliability coefficient of correlation is in computing the *predictive efficiency of a test*. Suppose we use the following formula: $E = (1 - \sqrt{1 - r^2}) 100$. By multiplying by 100 we change the answer into percentage of efficiency. Let us take a correlation coefficient and substitute it in the formula. Let r equal .80. Then

$$E = (1 - \sqrt{1 - .64}) 100 = 40 \text{ per cent efficient}$$

(see page 32). It is amazing how inefficient our best tests are when measured by this accurate formula. Even our best tests are only 68 per cent efficient, while those with lower reliability are correspondingly less efficient.

Validity

The validity of a test is usually obtained by correlating it with some criterion which indicates the more certain presence of what the test measures, or with other proved tests of the same trait. In our text we have mentioned the correlations of intelligence tests with success in life, and of group tests with individual tests such as the Stanford-Binet. The tests of the Army Air Force were correlated with success in flying, the Minnesota Mechanical Assembly Test with success of junior high students in a course in mechanics, and the Minnesota Clerical Test with the success of stenographers. During the First World War the scores on Army Alpha correlated .50 to .70 with officers' estimates of the success of their men. Finally, inventories of neuroticism have been correlated with other inventories and with the presence or absence of neurotic symptoms as discovered in a clinic. In a variety of correlations, indications are secured which point to the measurement by the test of those traits which it is attempting to measure.

Prognosis

Prediction is one of the most sought-after outcomes of testing. With what confidence can we predict from the present I.Q. of a child what I.Q. the child will have 3 years hence? Will this person who scores high

on the Minnesota Clerical Test be a success in stenography? Will that person who scores high on our battery of Air Force tests really get his wings? Whatever the answers are, they are determined by correlations. One of the questions frequently asked in school is, "Will this girl be a success in studying a foreign language?" In answering this question a prognostic test of language ability is given to a large group of students, their subsequent marks in a language are collected, and then a correlation is computed between the capacity as measured by the test and the success as estimated by the teacher or by an achievement test. Does an intelligence test prophesy subsequent college marks better than the high school records? In such a case correlations are computed between the test scores and college marks and between high school marks and college marks and an answer given in terms of the coefficient of correlation. Thus we say r between high school marks and college marks averages about .55 to .60, and between intelligence tests and college marks about .50 to .55.

The point is that once these relations are determined we can use the scores obtained at an earlier date to predict what persons will do at a later date. For example, those with top scores in the tests for aviators succeeded in flying in over 80 per cent of the cases, those with the lowest scores succeeded in less than 20 per cent of the cases.

Test Construction

Already under validation we have indicated that test items must correlate with the selected criteria. Ideally, test items would have a substantial correlation with the criterion and a low correlation with other items of the test. On the other hand for purposes of consistency each new item added to a test must correlate somewhat (at least .30) with the test as a whole. You will remember also that the computation of the amount of g which a test contains is determined by correlation. It is thus evident that every aspect of test construction is in some manner related to correlation. It thus may be truly said that *a test is known by its correlations*.

SAMPLING—STANDARD ERROR OF THE MEAN AND OF THE STANDARD DEVIATION

Every single statistic has a standard error of measurement, and in every case the interpretation is just like the $S.E._{meas.}$ which has been demonstrated. Let us take the mean as an illustration. If we wished to secure the mean height of college 18-year-olds we would draw from those available 100 cases at random. We would compute the mean. It would be clear to us that this was only a sample and that its relation to the whole would depend upon (1) the number of cases, and

(2) the size of the standard deviation. The true mean of height of 18-year-old college boys could be had by drawing out *all* this population and computing the mean. However this is not necessary, for the formula $S.E._{mean}$ gives us a clear indication of the limits between which the true mean would fall. $S.E._{mean} = S.D./\sqrt{N-1}$. Let us assume that the mean we computed for 100 cases is 68 inches, with a σ of 2.6 inches. Then $S.E._{mean} = 2.6/\sqrt{100-1} = .26$. We can now say the chances are 68 out of 100 that the true mean lies within ± 1 S.E., *i.e.*, between 67.74 and 68.26 inches; that the chances are 95 to 5 that the true mean lies between 67.48 and 68.52; and finally, that the chances are more than 99 out of 100 that the true mean lies between 67.22 and 68.78 inches. In like manner is interpreted the standard error of the standard deviation $S.E._s = S.D./\sqrt{2(N-1)}$.

SUMMARY

Statistical method is used in the construction and interpretation of tests and in their application. Scores on tests of any kind are arranged in tables of distribution from which much can be learned by inspection. Measures of central tendency—mean, median, and mode—may then be computed. The most important of these is the mean. It, however, is greatly influenced by extreme cases. When these extreme cases are accidental or not truly representative the median increases in importance. Measures of dispersion or scatter state quantitatively the amount of clustering of the scores around the central tendency. The standard deviation is the most reliable of the measures of dispersion. The semi-interquartile range (Q), the average deviation (A.D.), and the probable error (P.E.) are other measures of dispersion. The standard deviation may be used in computing T-scores or standard scores. These scores are better than percentile scores because they are based on the true distribution of scores.

The coefficient of correlation indicates the average degree of resemblance found between two traits in the same group of individuals when each individual is measured twice. Two procedures for computation, the Pearson product-moment method and Spearman method of rank differences, are introduced. The uses of this coefficient are legion. In computing prognosis, reliabilities, and validities of tests this coefficient is indispensable. Its substitution in formulas to denote the reliability of scores and the efficiency of tests adds greatly to our understanding of these terms.

Running through our whole treatment is the concept of sampling. One measure of an individual is merely a sample of his performance, not the true measure. The mean of a small random sample of any population

is just one of the possible means which other samples similarly drawn would show. Fortunately, from a single score, coefficient of correlation, measure of central tendency, measure of dispersion, ranges within which the true score lies may be calculated and the level of confidence in each range indicated. No concept in statistics helps more in the interpretation of these scores than that of sampling.

QUESTIONS AND EXERCISES

1. Distinguish between the mean, median, and mode. Which measure is most influenced by extreme cases? Why?

2. The following are actual scores made on a test of word knowledge by college students, the highest possible score being 150:

| | | |
|-----|-----|-----|
| 105 | 126 | 103 |
| 110 | 94 | 124 |
| 125 | 115 | 65 |
| 96 | 112 | 106 |
| 124 | 107 | 131 |
| 118 | 107 | 126 |
| 118 | 114 | 88 |
| 116 | 118 | |
| 117 | 119 | |
| 104 | 139 | |
| 105 | 96 | |
| 108 | 108 | |
| 107 | 119 | |
| 123 | 122 | |
| 61 | 112 | |
| 116 | 129 | |

Make a table of distribution from these data, using a convenient interval of 5 or 7. Define accurately the beginning and end of each step. *Make all computations from this table of distribution.*

3. From the above table, compute (a) the median and the 40th, 25th, and 75th percentiles; (b) the mean from the assumed mean; (c) the standard deviation and Q.

4. Suppose this table were a representative sample of a defined population; how would you calculate norms?

5. Compute several T-scores from this distribution. Why is it that a T-score is more accurate than a percentile score?

6. In the accompanying table are 25 pairs of scores: X (health knowledge) and Y (socioeconomic level).

| | Health knowl- edge, X | Socioeconomic level, Y |
|----|----------------------------|-----------------------------|
| 1 | 53 | 15 |
| 2 | 50 | 31 |
| 3 | 48 | 14 |
| 4 | 50 | 14 |
| 5 | 49 | 14 |
| 6 | 49 | 24 |
| 7 | 48 | 4 |
| 8 | 52 | 21 |
| 9 | 49 | 16 |
| 10 | 46 | 7 |
| 11 | 51 | 14 |
| 12 | 49 | 14 |
| 13 | 48 | 7 |
| 14 | 52 | 13 |
| 15 | 47 | 19 |
| 16 | 48 | 7 |
| 17 | 45 | 11 |
| 18 | 40 | 13 |
| 19 | 51 | 16 |
| 20 | 42 | 8 |
| 21 | 45 | 14 |
| 22 | 45 | 13 |
| 23 | 46 | 12 |
| 24 | 43 | 10 |
| 25 | 45 | 13 |

a. Compute r (1) by the Pearson method, (2) by the Spearman method of rank differences.

b. Interpret this coefficient as to its size and as to its reliability (apply here the standard error of r).

c. How does the problem of sampling enter into your interpretation of r ?

7. a. How can the standard error of measurement be used to interpret the meaning of a score?

b. Given a reliability coefficient of .90, a mean of 50, and an S.D. of 10

(S.D. the same on each form), if a subject scored 27, within what limits would his true score lie? State the level of confidence in each case.

8. Given a mean of 63, an S.D. of 10, and an N of 121, within what limits would the true mean lie? How does sampling enter into the interpretation?

BIBLIOGRAPHY

GARRETT, HENRY E.: *Statistics in Psychology and Education*, 3d ed. New York: Longmans, Green & Co., Inc., 1947.

GUILFORD, J. P.: *Fundamental Statistics in Psychology and Education*, 2d ed.

New York: McGraw-Hill Book Company, Inc., 1950.

WALKER, HELEN M.: *Elementary Statistical Methods*. New York: Henry Holt and Company, Inc., 1943.

Index

A

- Aamodt, Geneva P., 247
- Abbott, Allan, 173-174
- Achievement-test batteries, 79-93
 - development of, 79-82
 - evaluation of, 87-90
 - geography, 189-191
 - language, 146-149
 - literature, 152-156
 - mathematics, 226-228
 - reading, 96-97
 - science, 250-252
 - social, 186-189
 - spelling, 122-123
 - types of, 82-87
 - uses of, 90-92
- Achievement tests, characteristics, 9-10
 - constructing, 40-66
 - essay-type questions, 41-43, 57-63
 - organization and arrangement, 56-57
 - short-answer questions based on, recall, 43-47
 - recognition, 47-55
 - short-answer tests, higher mental processes, 55
 - validity, 15-21
- Adkins, D. C., 442-446
- Administering of tests, 72
- Administrability of tests, 34-35
- Algebra, and geometry, prognostic tests of, 240-241
 - objectives in teaching of, 232-233
 - tests of, 234-237
 - prognostic, 241
- Allen, Mildred M., 39
- American Council Civics and Government Test, 195
- Anderson, Roy N., 287
- Anderson, Theresa W., 339
- Anderson, W. N., 120, 142
- Andrew, Dorothy M., 274
- Appreciation of literature, measurement of, 172-178
- Aptitude tests, 22, 24
- Aptitude tests, for art, 299-306
 - for mechanics, 317-329
 - for music, 288-294
- Arithmetic, survey batteries, 226-228
 - tests of, 226-232
 - diagnostic, 229-232
 - separate, 228-229
- Army Alpha and Army Beta, 379-381
- Arthur, Grace, 371, 376
- Arthur's Point Scale of Performance Tests, 371
- Arts, achievement, 306-307
 - capacity in, 299-306
 - measurement of, 298-307
 - objectives in the teaching of, 298-299
- Ashbaugh, E. J., 124, 142
- Aspects of personality (test), 475-477
 - construction and scoring, 476-477
 - three dimensions of, 475
- Attitude scale, construction of, 451
- Attitudes, changes in, 460
 - definition of, 447-448
 - description of, 449-450
 - learning of, 448-449
 - measurement of, 450-460
 - in social sciences, tests of, 201-202
 - uses of scales, 460-462
- Ayres, L. P., 123, 142, 143
- Ayres Measuring Scale for Handwriting, 130-131
- Ayres Spelling Scale, 123-124

B

- Babcock, Harriet, 334
- Ball, Rachel S., 491, 494-495
- Bare, T. H., 463
- Barrett, Dorothy M., 287
- Barrett-Ryan-Schrammel English Test, 158
- Batteries of fundamentals, 83-87
- Becker, Ida S., 246
- Beliefs on social issues, test of, 455-456
- Bell, Hugh M., 471, 494
- Bell Adjustment Inventory, 471-473
 - divisions of, 471

- Bell Adjustment Inventory, interpretation of, 472
 validity of, 472-473
- Bennett, George K., 334
- Bernreuter, Robert G., 376
- Bernreuter Personality Inventory, 468-471
 nature and construction, 468-469
 scoring of, 469
 validity of, 470-471
- Betts, E. A., 141
- Binet, Alfred, 10, 354
- Bingham, Walter Van Dyke, 39, 274, 287, 316, 334
- Biology tests, 255-258
- Bixler, Harold E., 161
- Bixler High School Spelling Test, 161
- Blackstone, E. G., 287
- Blaisdell Instructional Tests in Biology, 262
- Bloom, Benjamin S., 39
- Bogardus, E. L., 464
- Bogardus Scale of Social Distance, 453
- Bookkeeping tests, 280-284
 list of, 283-284
 United-NOMA Business Entrance Tests, 282-283
- Bookwalter, Karl W., 350
- Bovard, John F., 335, 340, 349
- Brace, David K., 341, 349
- Brace Scale of Motor Ability Tests, 341
- Breed, F. S., 142
- Broom, M. E., 142, 334
- Brown, Clara M., 314, 315
- Brownell, Clifford Lee, 340, 349
- Brownell, W. A., 232
- Brownell's Posture Silhouette Scale, 340
- Bruner, Herbert B., 463
- Buchanan, Milton A., 223
- Buros, Oscar K., 69, 205, 223, 233, 246, 283, 287, 322, 419, 494
- Burt, Harold E., 417, 419
- Burt Agricultural Interest Test, 438
- Business content tests, 284-285
 list of, 286
- Business education, measurement of, 273-287
 objectives in, 273
 tests, bookkeeping, 280-284
 clerical, 274-280
 content, 284-286
- Business Fundamentals and General Information Test of United-NOMA Business Entrance Tests, 285
- Buswell, G. T., 246
- Buswell-John Diagnostic Test for Fundamental Processes in Arithmetic, 230-231
- California Achievement Tests, 85-89, 94, 122, 149, 227-228, 231-232
- California Aptitude Tests for Occupations (Roeder and Graham), 326-328
- California Group Functional Test (Stolz), 338
- California Intelligence Test, 394-396
- California Test of Personality, 473-475
 dimensions of, 473-474
 inventories for all grades, 473
 validity of, 474-475
- Canning, L. B., 446
- Cardiovascular tests, 336-338
- Carey, Stephen M., 464
- Carroll, Herbert A., 175, 182, 333
- Carroll Prose Appreciation Test, 175-178
- Carter, H. D., 446
- Carter, Ralph C., 59, 65
- Cattell, J. McKeen, 353
- Chase, Stuart, 273
- Chave, E. J., 451, 464
- Chemistry tests, 258-260
- Cheydleur, F. D., 223
- Civics tests, 195
- Civilian occupations, AGCT scores, 414-416
- Clark, R. S., 409
- Clark, Willis W., 94, 142
- Classroom tests, constructing, 40-57
 essay-type questions, 41-43, 57-63
 higher mental processes, 55
 organization and arrangement, 56-57
 matching, 52-55
 multiple-choice, 47-49
 sentence-completion, 46-47
 short-answer questions, 43-55
 true-or-false, 49-52
- Cleeton, Glen U., 440, 446
- Cleeton Vocational Interest Inventory, 430-431
- Clerical achievement tests, 276-280
- Clerical content tests, 284-286
- Clerical tests, achievement, 276-278
 aptitudes, 274-276
- Cole, Luella, 200
- College success prediction and intelligence tests, 410-412
- Columbia Research Bureau Algebra Tests, 234-235
- Commercial education survey test, 278-279
- Compass Diagnostic Tests in Arithmetic, 229-230
- Complete batteries of achievement tests, 82-83
- Conard, Edith U., 143

Conard Manuscript Writing Standards,
131-133

Concepts used in social sciences (Pressey
test), 200-201

Cook, Walter W., 403-404

Cooke, Dennis H., 246

Cooperative Algebra Test, 235-237

Cooperative American History Test, 192-
193

Cooperative Biology Test, 255-257

Cooperative Chemistry Test, 258-260

Cooperative Economics Test, 194

Cooperative English Test, effectiveness of
expression, 162-163

mechanics of expression, 158

organization, 163-164

reading comprehension, 167-170

spelling, 162

Cooperative French Test, 209-210

Cooperative General Achievement Tests,
196-197

Cooperative German Test, 214-216

Cooperative Latin Test, 217-219

Cooperative Literary Acquaintance Test,
178

Cooperative Mathematics Test for Grades
7, 8, and 9, 228-229

Cooperative Modern European History
Test, 193

Cooperative Physics Test, 260-261

Cooperative Plane Geometry Test, 238-240

Cooperative Science Test for Grades 7, 8,
and 9, 252-254

Cooperative Social Studies Test for Grades
7, 8, and 9, 189, 195-196

Cooperative Spanish Tests, 211-213

Coordinated Scales of Attainment, 78, 94,
153, 186-187, 228, 250-251

Correlation, coefficient of, 509-518
interpretation of, 514-515

Pearson product-moment method, 510-
512

Spearman rank-difference method, 512-
514

uses of, 515-518

Courtis, S. A., 14

Courtis Research Tests in Arithmetic, 15

Cozens, Frederick W., 335, 340, 342, 349, 350

Crawford, John Edmund, 324

Cronbach, Lee, Jr., 13, 15, 26, 39, 65, 93,
419, 494

Cruikshank, Ruth M., 334

Cubberley, Hazel J., 342, 349

Cureton, Thomas K., 350

Cureton, Thomas K., Jr., 350

Curtis, Dwight K., 271

D

Darley, J. G., 494

Dashiell, J. F., 448

Daugherty, M. L., 143

Davis, G., 142

Davis, H., 404, 419

Davis, Ira C., 271

Detroit Mechanical Aptitude Examination
for Girls, 318

Dewey, B., 323

Dewey, John, 424

Diagnostic Test for Fundamental Processes
in Arithmetic (Buswell and John),
230-231

Diamond, Leon N., 271

Dickson, V. E., 404

Differential Aptitude Tests, 328-329

Drake, Raleigh M., 333

Duran, June C., 334

Durrell, Donald D., 141, 419

Durrell Analysis of Reading Difficulty,
115-117

E

Economics tests, 194

Economy, 37

Edgren, H. D., 350

Educational guidance and intelligence tests,
407-409

E.R.C. Stenographic Aptitude Test, 275-
276

Eldridge, R. C., 120, 142

Ellingson, Mark, 435, 446

Elliott, Edward C., 3, 39, 43

Ellis, Albert, 482, 494

Emerson, Marion Rines, 334

Engle-Stenquist Home Economics Test,
312-313

English composition, scales of, 164-167

Hillegas, 165

Hudelson, 165-167

Lewis, 165

Nassau County, 165

Van Wagenen, 165

English-usage tests, 158-160

English vocabulary tests, Cooperative, 171
Inglis, 171

Equal-appearing units, 451

Espenchade, Anna, 350

Essay examination, weakness of, 3

Essay-type questions or examinations, 41-
43, 57-63

causes of unreliability, 41-43

improvement, of questions, 59-60
of scoring, 61-63

value of, 58-59

- Examination, in bookkeeping and accounting, 280-282
in plane geometry, 238
- F
- Farnsworth, P. R., 289, 292, 333
Faubion, Richard, 322
Faulkner, Ray, 333
Feder, Daniel D., 446
Feeble-minded, interest in, and intelligence tests, 354-355
Filer and O'Rourke Rating Scales, 485-486, 494
Filing test, United-NOMA Business Entrance Tests, 284
Fine arts and manual arts, measurement of, 288-334
 arts, fine, 298-307
 manual, 307-329
 music, 288-298
 objectives, 295, 298-299, 308-309
Flanagan, J. C., 271, 468, 469, 494
Foran, Thomas G., 94, 142
Foreign languages, measurement of, 207-224
 objectives in teaching, 207-208
 tests, French, 208-211
 German, 214-217
 Latin, 217-220
 Spanish, 211-214
Forrest, Ruth, 377
Foster, J. G., 376
Foster, Josephine C., 376
Fransden, Arden, 446
Franseen Diagnostic Tests in Language, 151-152
Freeman, F. N., 135, 143, 361, 376
Freeman Chart for Diagnosing Faults in Handwriting, 134-135
French, Esther, 343, 350
French tests, 208-211
 lists of, 210-211
Frequency of occurrence, check on validity, 16
- H
- Frœhlich, Gustav J., 411, 419
Frutchey, Fred P., 271
Fryer, Douglas, 419, 438, 445
- G
- Gage, N. L., 13, 39, 66, 182, 446, 447, 453, 463
Galton, Sir Francis, 511
Gardner, Iva Cox, 461, 464
Garretson, O. K., 441
Garrett, Henry E., 28, 29, 39, 369, 515, 521
Gates, A. I., 125, 141, 142, 344
Gates Tests of Reading, 104-106, 109-111
Gates-Strang Health Knowledge Test, 344-345
General science, tests of, 252-255
Geography tests, 189-191
Geometry tests, 237-240
German tests, 214-217
 list of, 216-217
Gerberich, J. Raymond, 93, 141, 181, 203, 223, 246, 271, 287
Gist, A. S., 95
Glenn-Greenberg Instructional Tests in General Science, 262
Glenn-Obourn Instructional Tests in Physics, 262
Glenn-Welton Instructional Tests in Chemistry, 262
Goddard, Eunice R., 224
Goddard, Henry H., 10, 354
Goodenough, Florence L., 13, 376
Goodenough "Drawing a Man" Scale, 371-372
Goodman, Charles H., 376
Grade equivalent, 81-82
Gray, C. T., 137, 143
 Standard Score Card for Measuring Handwriting, 137
Gray, H. A., 271
Gray, W. S., 142
Gray's Oral Reading Test, 106-108
 individual record sheet, 118
Gray-Votaw-Rogers General Achievement Tests, 87, 94
Greene, Edward B., 333, 445, 494
Greene, Harry A., 93, 141, 181, 205, 223, 246, 271, 287
Grice Generalized Attitude Scale, 454
Grover, C. C., 246
Guidance measurement, 9
Guilford, J. P., 39, 494, 521
Guilford, R. B., 494
Guttman, L., 39
- H
- Haggerty-Olson-Wickman Behavior Rating Schedules, 11, 484-485, 488-490
Hagman, E. Patricia, 335, 340, 349
Handwriting, 127-138
 aims and objectives in teaching, 127-128
 diagnosis of, 134-136
 measurement of, 128-134
 practice exercises, 136-138
Handwriting Scale (E. L. Thorndike), 6
Handwriting score card, 137
Harrell, Willard, 322
Harrison, M. Lucile, 141
Hartley, Eugene L., 456, 457

- Hartley Picture Attitude Test toward Negroes, 456-459
- Hartog, Sir Philip, 4, 41
- Harvard Step Test, 337
- Hathaway, S. R., 470, 494
- Hauch, Edward F., 223
- Hawkes, Herbert E., 65, 181, 223, 271
- Health education, list of tests in, 347-348
- Health information tests, 343-345
- Health Inventory for High School Students (Neher), 345-347
- Health practices, 345-347
- Henmon, V. A. C., 221, 223
- Herring, John P., 376
- Hesler, Russell J., 287
- Hiett Stenography Test, 277
- Higher mental processes, tests of, 55
- Hilghsmith, J. A., 333
- Hildreth, Gertrude, 94, 142, 371, 376
- Hillbrand, E. K., 295
- Hillegas Scale for Measurement of Quality in English Composition, 165
- Hinckley, E. D., 464
- Hinckley Attitude Scale toward the Negro, 454
- History tests, 192-194
- Hoff, A. G., 271
- Home economics, measurement in, 312-316
- rating scales and check lists, 314-316
- tests for high school, 313-315
- Home-mechanics tests, 310-311
- Homogeneous grouping and intelligence tests, 409-410
- Horn, Ernest, 17, 39, 120, 142
- Horne, E. Porter, 449, 464
- Horning, S. D., 322, 334
- Horowitz, E. L., 464
- Howland, Amy R., 350
- Hoyer, Louis P., 443
- Hubbard, R. M., 442, 446
- Hudelson, Earl, 161
- Hudelson Typical Composition Ability Scale, 166-167
- Hunter, E. C., 459, 483
- Hunter Test of Social Attitudes, 459
- I
- Indiana Tests of Home Economics, 313-314
- Individual differences, 353-354
- in intelligence in same grade, 403-404
- I.Q. (intelligence quotient), 358-361, 383
- characteristics of, 360-361
- Intelligence tests, and beginning reading, 405
- and definition of feeble-minded, 411-412
- Intelligence tests, and election of high-school subjects, 405-407
- group, 378-419
- development of, 378-381
- types of, 381-390
- for grades 1 through 3, 391-396
- for grades 4 through 8, 396-400
- for high school, 400-403
- for kindergarten and first grade, 390-391
- uses of, 403-418
- individual, 353-377
- description of, 355-368
- development of, 353-355
- general nature of, 18
- and meaning of intelligence, 372-374
- performance tests, 368-372
- validity of, 21-24
- Interest and achievement, 441-442
- Interest measurement, 423-446
- Interests, characteristics of, 423-424
- correlation of, with achievement, 439
- through information, 435-439
- tests of, validity of, 438-439
- inventories, 426-435
- list of, 436-437
- uses of, 439-441
- methods of discovering, 424-426
- in relation to other traits, 441-444
- Interpretation and comparability of tests, 35-37
- Interpreting and using results of tests, 73-79
- Iowa Every-pupil Tests of Basic Skills, 84-85, 97, 148, 190-191
- Iowa Language Abilities Test, 149-151
- Iowa Silent Reading Tests, 17, 111-113
- Iowa Spelling Scales, 124
- Italian tests, 217
- J
- Jarvie, L. L., 435, 446, 470, 494
- John, Lenore, 246
- Johns, A. A., 470, 494
- Johnson, Guy B., 333
- Jones, W. Franklin, 120, 142
- Jordan, Arthur M., 23, 39, 267, 406-408, 419, 423-424, 434, 445, 478, 486, 510
- Jorgensen, Albert N., 93, 141, 181, 205, 223, 246, 271, 287
- Judgment of experienced observers, check on validity of, 17-18
- Jurgensen, Clifford, E., 287
- K
- Karnes, M. Ray, 65
- Karpovich, Peter V., 350
- Katz, S. E., 39

- Kaulfers, Walter Vincent, 220
 Kefauver, Grayson N., 419
 Kelley, Truman L., 31, 39, 58, 63, 65, 80, 94, 184, 206
 Kelly, Ida B., 464
 Kelly-Moore Test of Concepts in the Social Studies, 200
 Keniston, Hayward, 223
 Kent, Grace H., 361, 376
 Kilby, Richard W., 142
 King, W. A., 95
 Kintner, Madaline, 334
 Klugman, Samuel F., 287
 Knauber, Almer Jordan, 306, 334
 Knauber Art Ability Test, 306-307
 Knuth, William E., 333
 Koos, L. V., 128, 419
 Kopel, David, 142
 Kornhouser, A. W., 441, 446
 Krey, A. C., 58, 63, 65, 184, 206
 Krugman, M., 362
 Kuder, G. Frederic, 25, 29, 39, 425, 446
 Kuder Preference Record, 431-433
 Kuhlmann, F., 376
 Kuhlmann-Anderson Intelligence Tests, 384-387
 reliability of, 386-387
 validity of, 385-386
 Kwalwasser Test of Musical Information and Appreciation, 297-298
 Kwalwasser-Dykema Music Tests, 292-293
 Kwalwasser-Ruch Test of Musical Accomplishment, 296
- L
- Landis, Carney, 39, 470, 494
 Language, aims and objectives of teaching, 144-145
 and literature, measurement of, 144-182
 lists of tests in, elementary schools, 155
 secondary schools, 179
 tests in, elementary schools, 145-152
 secondary schools, 156-172
 written, tests in, 145-152
 diagnostic, 151-152
 separate, 149-151
 (See also Literature)
 La Porte, William L., 335
 Larson, Leonard, 350
 Latin tests, 217-220
 list of, 220
 Leamer, Emery W., 143
 Lectures, effect on attitudes, 461
 Lee, Doris May, 142, 247
 Lee, Edwin A., 433
 Lee, J. Murray, 142, 247
 Lee-Thorpe Occupational Interest Inventory, 433-434
 Lenz, Theodore F., 463
 Leonard, Ruth, 322, 334
 Lewerenz, Alfred S., 304, 334, 463
 Lewerenz Tests in Fundamental Abilities of Visual Arts, 304-306
 Lewis English Composition Scales, 165
 Lide, Edwin S., 238, 246
 Likert, Rensis, 463, 464
 Lind, Christine, 94
 Linden, Arthur V., 463
 Lindgren, Henry C., 434, 446
 Lindquist, E. F., 13, 65, 181, 206, 223, 246, 271
 Literary acquaintance (secondary school), tests of, 178-180
 Literary appreciation, tests of, 172-178
 Literature, and language (see Language)
 tests of, elementary schools, 152-156
 secondary schools, 172-180
 Logassa, Hannah, 182
 Longstaff, Howard P., 274
 Loutit, C. M., 480
 Loyes, Edmund, 94
- M
- McAdory, Margaret, 301
 McAdory Art Test, 301-304
 MacBroom, Maud, 17, 39
 McCall, William A., 7, 507
 McCall, William C., 446
 McCloy, C. H., 341, 350
 McCoy, Martha J., 182
 McHale, Kathryn, 438, 446
 McHale Vocational Interest Test for College Women, 438
 Machine calculation, United-NOMA Business Entrance Tests, 284
 MacMurray, Donald, 377
 MacQuarrie, T. W., 321
 MacQuarrie Tests for Mechanical Ability, 318, 320-323
 Madsen, I. N., 405, 419
 Maller Case Inventory, 477-478
 construction and validity, 477-478
 types of scores, 477
 Mann, C. R., 65, 181, 223, 271
 Manual arts, 307-310
 objectives in teaching of, 308-309
 tests of, 309-310
 Manuel, H. T., 25
 Matching tests, 52-55
 Mathematics, measurement of, 225-247
 objectives in teaching of, algebra, 232-233
 arithmetic, 225-226

- Mathematics, objectives in teaching of, geometry, 238
 tests, in elementary schools, 225-232
 list of, 242-245
 in secondary schools, 232-241
- Maurer, Katharine M., 21
- Mean, arithmetic average, 503-505
- Mean deviation, 507
- Measurement of intelligence (*see* Intelligence tests)
- Measuring of mental traits, difficulties in, 4-7
- Measuring instruments, administrability, 34-35
 characteristics of, 14-39
 economy, 37
 interpretation and comparability, 35-37
 reliability, 26-33
 validity, 14-26
- Mechanical-ability tests, assembly and performance, 318-323
 information, 317-318
 paper-and-pencil, 323-329
- Mechanical aptitude and ability, testing procedures, 316-329
 information about mechanical ability, 317-318
 mechanical assembly tests, 318-323
 paper-and-pencil tests, 323-329
 processes analyzed into elements, 317
- Mechanical Aptitude Test of United States Army, 438
- Mechanical interest test, 438
- Median, 501-502
- Meier, Norman C., 334
- Meier-Seashore Art Judgment Test, 299-301
- Mellenbruch, Paul L., 325
- Mellenbruch Mechanical Aptitude Test for Men and Women, 324-326
- Mental-age scales, 355-363
- Merrill, Maude A., 39, 357, 359, 363, 377, 384
- Metropolitan Achievement Tests, 82-83,
 91, 94, 98-103, 122, 146-147, 153-154,
 187-189, 226-228, 250-252
- Metropolitan Reading Readiness Test, 98-100, 103
- Micheels, W. J., 65
- Michigan Pulse Rate Test for Physical Fitness, 337
- Miller, Augustus T., 337, 350
- Minard, Ralph D., 459, 460, 464
- Minard Test of Racial Attitudes, 459-460
- Minnesota Check List for Food Preparation (Brown), 314-315
- Minnesota Food Score Cards (Brown), 315-316
- Minnesota Mechanical Assembly Test, 318-320
- Minnesota Paper Form Board Test, Revised, 323-324
- Minnesota Vocational Test for Clerical Workers, 274-275
- Mitchell, Mildred B., 377
- Mode, 505
- Monroe, Marion, 141
- Monroe, Walter S., 59, 65, 419
- Moody, Caesar B., 389-390
- Morehouse, Lawrence E., 337, 350
- Morgan, B. Q., 223
- Morgan, W. J., 334
- Morrison-McCall Spelling Scale, 124-125
- Morrow, Robert S., 287
- Mosher, Raymond M., 295
- Mosher Test of Individual Singing, 295-296
- Motor coordination tests, 341
- Moving pictures, effect on attitudes, 461
- Multiple-choice tests, construction of, 47-49
- Murphy, Gardner, 463
- Mursell, James L., 290, 333
- Music tests, 288-298
 objectives of, 295
- Musical aptitude, measurement of, 288-294
- Musical Aptitude Test (Whistler and Thorpe), 293-294
- Musical information, appreciation, and achievement, 294-298
- N
- Nash-Van Duzee Industrial Arts Tests, 309-310
- Neher, Gerwin Charles, 345
- Neilson, N. P., 342, 349
- Netzer, Royal F., 146
- Newcomb, T. M., 463
- Newkirk, Louis V., 311, 334
- Newkirk-Stoddard Home Mechanics Test, 310-311
- Newman, Horatio H., 360
- Noll, Victor H., 271
- Norms, local, 36
- Noyes, E. S., 62, 66
- O
- Objectives in education, 4
- Odell, C. W., 182, 223, 246, 271
- Oral English, 145-146
- Orleans, Jacob S., 66
- Orleans, Joseph B., 247
- Orleans Algebra Prognosis Test, 241
- Organization and arrangement of tests, 56-57

- O'Rourke Mechanical Aptitude Test, 323, 332
 Otis, Arthur S., 380

P

- Paterson, Donald G., 287, 319, 320, 334
 Pearson, John M., 246
 Pearson, Karl, 511
 Percentiles, 501-503
 Performance tests of intelligence, 368-372
 Perry, Fay V., 334
 Perry, Winona M., 247
 Personality inventories, measurement, of
 attitudes, 447-464
 of interest, 423-446
 of personality traits, 465-495
 Personality rating scale for preschool children, 491
 Personality traits, measurement of, 465-495
 rating scales, 483-491
 self-inventories, difficulties with, 466-468
 types of, 468-482
 validity of personality inventories, 482-483
 Peters, Emma, 223
 Peters, F., 464
 Peterson, Joseph, 376
 Peterson, Ruth, 461, 463
 Physical education, achievement tests, 342-343
 and health, measurement of, 335-350
 objectives in, 335-336
 rating scales, 348
 tests, of health information, 343-348
 of physical capacities, 336-342
 Physics tests, 260-262
 Pintner, Rudolph, 355, 371, 376, 409, 412, 419
 Pintner General Ability Tests, 381-383
 Pintner Intelligence Tests, Intermediate, Advanced, 383
 Pintner-Cunningham Primary Test, 382
 Pintner-Durost Elementary Intelligence Test, 383, 392-393
 Pintner-Paterson Scale of Performance Tests, 369-371
 Piper, A. H., 247
 Plan for testing program, 70-71
 Poetry, exercises in judging, 173-174
 Point scales of intelligence, 363-368
 Pooley, Robert, 172, 182
 Porteus, S. D., 376
 Powers, S. R., 405, 419
 Pressey, L. C., 143
 Pressey, S. L., 143
 Pressey Diagnostic Tests in English Composition, 159-160
 Pressey Test of Concepts Used in the Social Sciences, 200-201
 Price, Roy A., 206
 Primary mental abilities, 387-390
 Probable error, 506
 Problems, skills, and procedures of testing in social science, 195-199
 Proctor, W. M., 407-408, 419
 Prognostic tests, 219-220
 Luria-Orleans Modern Language Prognosis Test, 219-220
 Orleans-Soloman Latin Prognostic Test, 219
 Symonds Foreign Language Prognostic Test, 219
 Psychological and logical analysis, check on validity of, 18-21
 Pullias, Earl V., 94
 Pyle, William H., 142

Q

- Q, semi-interquartile range, 506-507

R

- Racial attitudes, measurement of, 456-459
 Rating scales, 11, 483-491
 list of, 492
 samples of, 488-491
 types of, 484-488
 Read, James Morgan, 206
 Reading, 95-117
 objectives in teaching of, 95-96
 spelling, and handwriting, measurement of, 95-142
 tests of, in achievement batteries, 96-97
 diagnostic, 114-119
 oral, 117
 reading achievement, elementary school, 103-113
 high school, 113-114
 reading readiness, 97-103
 Thorndike-McCall, 7
 Reading comprehension (secondary school), 167-170
 Reading diagnosis, tests of, 114-119
 Reading tests, lists of, reading achievement, 113
 reading diagnosis, 119
 reading readiness, 101-102
 Ream Social Relation Test, 438
 Reliability, 26-33
 factors affecting, 29-32
 interpretation, 32-33
 methods for computing, 27-29
 Remmers, H. H., 13, 39, 66, 182, 446, 447, 453, 454, 463, 464

- Rhodes, E. C., 4, 41
 Richardson, M. W., 29, 39
 Rigg, Melvin G., 174, 182
 Measuring the Ability to Judge Poetry,
 174-175
 Rinsland, Henry D., 49, 53, 61
 Roberts, Catharine Ellis, 491, 494-495
 Roeber, Edward C., 434, 446
 Rogers, Frederick Rand, 339, 350
 Rogers Strength Test, 339
 Rogers Test of Personality Adjustment,
 479-481
 divisions of, 479
 types of tests, 479-480
 usefulness of, 480
 Rosanna, M., 464
 Ross, C. C., 13, 39, 43, 66, 182
 Ruch, G. M., 66, 80, 94, 223
 Ruch-Cosman Biology Test, 257-258
 Ruch-Popenoe General Science Test, 254-
 255
 Russell, D. H., 125, 142
- S
- Scates, Douglas E., 39
 Schlink, F. J., 273
 Schneider, E. C., 336, 350
 Schneider Test of Pulse Rate and Blood
 Pressure, 336-337
 Schneider, Gwendolen G., 287
 Schnell, Leroy N., 237
 Schoen, Max, 333
 Science, measurement of, 248-272
 aims and objectives, 248-249
 attitudes and interests, 266-267
 scientific thinking, 263-266
 tests, in elementary schools, 249-255
 in secondary schools, 255-262
 Science tests, instructional, 262
 list of, 268-270
 of understanding, 263-266
 Scientific attitudes and interests, 266-267
 Scientific thinking, 263-266
 Score cards in handwriting, 135-137
 Scoring tests, 72-73
 Scott, M. Gladys, 343, 350
 Seagoe, May V., 247
 Sealey, Glenn A., 66
 Seashore, Carl Emil, 288, 333
 Seashore's Measures of Musical Talent,
 289-292
 Seibert, Louise C., 224
 Selection of high-school subjects and intel-
 ligence, 405-407
 Self-inventories, 466-483
 difficulties in use of, 466-468
 Self-inventories, list of, 481
 types of, 468-482
 validity of, 482-483
 Sentence-completion tests, construction of,
 46-47
 Sentence-organization tests, 163-164
 Sentence-structure tests, 162-163
 Shanner, W. M., 25
 Sharpe, S. E., 377
 Short-answer tests, 43-55
 based on, recall, 43-47
 recognition, 47-55
 Shotwell, Anna M., 94, 141, 404
 Siceloff, Margaret McAdory, 301
 Simmons, Ernest P., 161
 Simple recall tests, construction of, 44-46
 Sims, Verner, 61, 66
 Sixteen spelling scales, 161
 Smith, Dora V., 144, 182
 Smith, Eugene R., 13, 18-21, 39, 55, 66, 173,
 182, 206, 263-264, 272, 446, 450, 455,
 463
 Smith, F. T., 463
 Social-science tests, list of, 203-205
 Social sciences, measurement of, 183-206
 objectives in teaching of, 184-186
 tests of social studies, elementary
 schools, 186-191
 secondary schools, 191-202
 Social terms, tests of, 199-201
 Social utility, check on validity of, 18
 Spache, George, 94
 Spanish tests, 211-214
 list of, 213-214
 Spearman, Carl, 368, 512
 Speer, G. S., 470, 495
 Spelling, 117-128
 objectives in teaching of, 121
 selection of word lists, 117-121
 tests of, elementary school, 121-125
 list of, 125
 secondary school, 160-162
 uses of, 125-127
 Spencer, Douglas, 425
 Stalnaker, John M., 62, 66
 Standard deviation, 505-506
 uses of, 507-508
 Standard error, of the mean, 518-519
 of measurement, 33
 of the standard deviation, 519
 Standard score, 81, 507
 Stanford Achievement Test, 82-83, 94, 122,
 147-148, 153, 189, 228, 250-251
 Stansbury, Edgar, 339
 Stanton, Hazel Martha, 291, 333
 Starch, Daniel, 3, 39, 43
 Starch, David, 120

- Statistical methods, 499-521
 assembling of data, 500-502
 central tendency, 502-505
 concepts, 499-500
 correlation, 509-515
 dispersion, 505-509
 sampling, 518-519
 uses of coefficient of correlation, 515-518
- Steadman, Robert F., 206
- Steinmetz, Harry C., 463
- Stenographic test, United-NOMA Business Entrance Tests, 277-278
- Stenographic tests, 275-278
 achievement, 276-278
 aptitude, 275-276
- Stenography and typewriting, list of tests in, 279-280
- Stern, Wilhelm, 372
- Stetson, F. L., 161
- Stewart, Naomi, 419
- Stoddard, George D., 223
- Stogdill, Emily, 470, 495
- Stolz, H. R., 338
- Stone, Clarence R., 142
- Stoy, E. G., 322, 334
- Strang, Ruth, 344, 350
- Strength tests, 338-341
- Strong, Edward K., 429, 446, 484
- Strong Vocational Interest Blank, 426-429
 history of, 426
 scoring, 427
 types of items, 427
 validation, 428-429
- Stutsman, Rachel, 376
- Super, Donald E., 13, 25, 442, 446, 494, 495
- Symonds, P. M., 182, 224, 246, 247, 441, 494

T

- T-score, advantages of, 81, 508-509
- Taylor, Katherine Van F., 446
- Teacher-made tests, organization of items, 56-57
- Terman, Lewis M., 23, 39, 80, 94, 354, 355, 357, 363, 376, 384
- Terman-Merrill Revision, 355-363
 evaluation, 361-363
 I.Q., 359-361
 mental age, 358-359
 principles of construction, 356-358
- Test of Critical Thinking in the Social Studies (Wrightstone), 197-199
- Testing program, 67-94
 administering and scoring of tests, 72-73
 interpretation of results, 73-79
 planning, 67-72
- Tests, administering of, 72

- Tests, administrability of, 34-35
 of foreign languages, evaluation of results, 221-222
 interpretation of results, 73-79
 teacher-made, organization of items, 56-57
 (See also specific names and subjects of tests)
- Thinking, methods of testing, 19-21
- Thomas, Minnie E., 470, 495
- Thorndike, Edward L., 120, 130, 133, 142, 143, 376, 405
- Thorndike Scale for Handwriting of Children, 130
- Thorpe, Louis P., 433
- Thurstone, L. L., 376, 388, 451, 452, 461, 463, 464
- Thurstone, Thelma Gwynn, 388
- Thurstone Attitude toward Communism Scale, 452
- Tidyman, W. F., 142
- Tiegs, Ernest W., 94, 141
- Tiffin, Joseph, 334
- Tonne, Herbert A., 273, 287
- Toops General Interest Test for Girls, 438
- Torgerson, T. L., 247
- Townsend, Agatha, 141, 206
- Trabue, M. R., 16, 173-174
- Trabue's Nassau County Scale of English Composition, 165
- Travers, Robert M. W., 66
- Traxler, Arthur E., 94, 182, 206, 224, 259, 446
- Trieb, Martin H., 342, 349
- Triggs, Frances Oralind, 433, 441, 446
- True-or-false tests, 49-52
- Turney, Austin H., 387
- Turse, Paul L., 287
- Turse-Durost Shorthand Achievement Test, 276-277
- Tyler, Ralph W., 13, 18-21, 39, 55, 66, 173, 182, 206, 263-264, 272, 446, 450, 455, 463
- Typing achievement tests, 278-280

U

- United-NOMA Business Entrance Tests, 284-285
- Units of measurement in education, 7-9

V

- Validity of tests, 14-26
 external, 22-24
 internal, 15-22
 recent trends in, 24-25

- Validity of tests, vitiating factors in, 25-26
Van Alstyne, Dorothy, 486, 495
Van Alstyne Rating Scales, 486
Vander Beke, George E., 223
Van Wagenen, M. J., 376
Van Wagenen English Composition Scales, 165
Vocabulary-load-of-interest inventories, 434-435
Vocabulary tests, 170-171
Vocational guidance and intelligence tests, 412-417
- W
- Walker, Helen M., 521
Webb, L. W., 94, 141, 404
Wechsler, David, 363, 364, 366, 376
Wechsler-Bellevue Intelligence Scale, 363-368
 adult intelligence, 364-366
 distinctive features, 367-368
 evaluation of, 366-367
 verbal and performance, 364
Weidemann, C. C., 59, 66
Wellman, Beth, 376
Wells, F. L., 373
Werner, Oscar H., 406
Wesley, Edgar Bruce, 206
Wesley Test, in political terms, 199
 in social terms, 199-200
West, Paul V., 129, 143
Whitford, W. G., 298
Wiedefeld-Walther Geography Test, 189-190
Winnetka Scale for Rating School Behavior, 490-491
Wissler, Clark, 377
Wittenborn, J. R., 446
Witty, Paul A., 142
Woodworth, R. S., 373, 423, 448, 465, 466
Woodworth Psychoneurotic Inventory, 466
Woodyard, Ella, 161, 301
Wolf, Henriette, 94
Wrightstone, J. Wayne, 142, 197, 464
Wrightstone Scale of Civic Beliefs, 201-202, 224
Wrightstone Test of Critical Thinking, 197-199
- Y
- Yerkes, Robert M., 376
Yoakum, Clarence S., 426
- Z
- Zapf, Rosalind M., 272

Form No. 3.

PSY, RES.L-1

**Bureau of Educational & Psychological
Research Library.**

The book is to be returned within
the date stamped last.

- 6 MAR 1961

5.9.62

11 JUN 1965

13.3.68

27.9.68

1.6.70

371.26
JOR

Form No. 4

BOOK CARD

Coll. No. 371.26 Accn. No. 592

Author Jordan, A. M.

Title Measurement in Educa
tion

| Date. | Issued to | Returned on |
|------------|-----------|-------------|
| 6 MAR 1961 | 3261 | 6 MAR 1961 |

371.26
JOR